Universidad Nacional de la Amazonía Peruana

Facultad de Ingeniería de Sistemas e informática

"MINERÍA DE DATOS PARA LA INTELIGENCIA DE NEGOCIOS"

INFORME PRÁCTICO DE SUFICIENCIA

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS E INFORMÁTICA

PRESENTADO POR EL BACHILLER:

EDSON VASQUEZ VALLES

IQUITOS – PERÚ

2015

INFORME TÉCNICO DEL EXAMEN DE SUFICIENCIA PREVIA ACTUALIZACIÓN ACADÉMICA APROBADO EN SUSTENTACIÓN PÚBLICA, POR EL JURADO EXAMINADOR, DESIGNADO POR EL DECANO DE LA FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA DE LA UNIVERSIDAD NACIONAL DE LA AMAZONÍA PERUANA.

ING. SAÚL FLORES NUNTA
Presidente

ING. CARLOS ALBERTO GARCÍA CORTEGANO
Primer Miembro

ING. FRANCISCO MIGUEL RUIZ HIDALGO
Segundo Miembro

Dedicatoria

Dedico este esfuerzo personal y este logro académico y profesional:

A mis Padres, Por darme la vida, creer en mí y apoyarme en todo momento, por sus consejos, sus valores, por la motivación constante que me ha permitido ser una persona de bien. Por los ejemplos de perseverancia y constancia que los caracteriza y que me ha infundado siempre, por el valor mostrado para salir adelante y por sus gran amor incondicional de Padres y Amigos a la vez.

A mi novia, por comprenderme y apoyarme siempre y en todo momento.

A mis compañeros de facultad, que nos apoyamos mutuamente en nuestra formación profesional y que compartieron esta carrera que por momentos parecía infinita.

A todos, espero no defraudarlos y contar siempre con su valioso apoyo sincero e incondicional.

Todo este trabajo ha sido posible gracias a ellos.

Agradecimiento a:

Dios, por darme la oportunidad de vivir y por estar conmigo en cada paso que doy, por fortalecer mi corazón e iluminar mi mente y por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía durante todo el periodo de estudio.

A nuestros maestros quienes nos han enseñado a ser mejores en la vida y a realizarnos profesionalmente.

A los compañeros de clases quienes nos acompañaron en esta trayectoria de aprendizaje y conocimiento.

A mis compañeros de trabajo de PETROPERU, por permitirme realizar este proyecto poniéndose a mi disposición y brindándome todas las facilidades desde el primer al último día en que así lo requerí.

En general quisiera agradecer a todas y cada una de las personas que han vivido conmigo en la realización de este trabajo.

Resumen

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 potencialmente útil de grandes bases de datos que puede ser utilizada como soporte para la toma de decisiones.

¿Cómo es exactamente que la minería de datos es capaz de extraer información importante que no se sabe que va a suceder después? la técnica que se utiliza para ejecutar estos hechos en la minería de datos se llama modelado de datos. El proceso de construcción de un modelo es algo que las personas han estado haciendo durante mucho tiempo, desde antes de la aparición de las computadoras. Lo que pasa con las computadoras, no es muy diferente de la forma en que las personas construían sus modelos. Las computadoras son cargadas con una gran cantidad de información sobre una variedad de situaciones donde se cuenta con una respuesta conocida y entonces el software de minería de datos en la computadora debe analizar a través de esos datos y determinar las características que se deben examinar por medio del modelo.

Al ser la minería de datos un método para extraer conocimiento útil mediante el análisis de los datos, ésta recurre a modelos que permitan encontrar relaciones, patrones o reglas inferidas previamente desconocidas. Los modelos empleados en la minería de datos son el descriptivo y el predictivo. Dentro de estos modelos se encuentran diferentes tareas específicas como: agrupamiento, reglas de asociación, clasificación, regresión, entre otras.

Al igual que estas tareas las redes neuronales también forman parte importante de la minería de datos ya que se basa en ésta para utilizarse en un gran número y variables de aplicación tanto comerciales como militares. La mayoría de estas aplicaciones consisten en realizar un reconocimiento de patrones, como: buscar un patrón en una serie de ejemplos, clasificar patrones, completar una señal a partir de valores parciales o reconstruir el patrón correcto partiendo de uno distorsionado.

Para mejorar la toma de decisiones existen infinidades de herramientas de Data Mining, estas herramientas nos ahorraran un arduo trabajo lo que permitirá generar modelos predictivos entre otras cosas de manera rápida, fácil y precisa.

Índice General

Índice de Figuras

Glosario de Términos y Abreviaciones

Término	Significado
Know-how	Se define como “saber cómo hacer algo pronto y bien hecho”. Se a los conocimientos preexistentes, no siempre académicos, que incluyen técnicas, información secreta, teorías e incluso datos privados como clientes o proveedores
Data warehouse	Es un almacén de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.
SAS	Sistema de Análisis Estadístico, es el nombre de un pionero en la inteligencia de negocio y una familia de software de bases de datos comercializados por la compañía SAS Institute Inc.
SEMMA	Es un acrónimo que significa Muestra, Explora, Modificar, Modelo, y Evaluar. Es una lista de pasos secuenciales desarrollados por SAS Institute Inc., uno de los mayores productores de estadísticas y software de inteligencia de negocio.
CRISP-DM	Cross-Industry Standard Process para la Minería de Datos
CBR	Razonamiento Basado en Casos
VLSI	Very Large Scale Integration, integración a escala muy grande.

I. Justificación

El desarrollo de las empresas modernas, en términos de competitividad y productividad, requiere de la construcción de arquitecturas empresariales que ayuden a gestionar las grandes cantidades de información que manejan durante años para que gerentes, ejecutivos, analistas, etc., puedan observar de manera gráfica e interactiva, los datos transformados en patrones históricos y así, poder tomar mejores decisiones de negocios a partir de los procesos que se analizan y se gestionan. Para estos propósitos es fundamental la transformación, planeamiento e implementación de grandes cantidades de información utilizando herramientas de Data Mining.

La inteligencia de negocio debe entenderse como un conjunto de estrategias que apoyan la toma de decisiones en las empresas, es por ello que la minería de datos encaja en este concepto ya que encamina la explotación eficiente de los datos mediante técnicas estadísticas y extracción de conocimiento procesable implícito en lo mismo.

En la actualidad, vivimos en una época en que la información es clave para obtener ventaja competitiva en el mundo de los negocios, por ello existe gran interés comercial por explotar los grandes volúmenes de información, pero no saben de qué forma se puede transformar toda esa información en conocimiento o sabiduría que apoye efectivamente la toma de decisiones, especialmente, a nivel gerencial. Una forma de llenar esos vacíos es usar los métodos y herramientas de la minería de datos apoyándose en la inteligencia de negocios, de tal manera, que los grandes volúmenes de información procesada sean usados adecuadamente.

Alrededor del mundo, se ha estimado el crecimiento de los datos almacenados en las bases de datos se duplican cada 18 meses mientras que la técnica de análisis de información no han tenido un desarrollo equivalente, dicho en otras palabras, la velocidad en que se almacena la información es muy superior a la velocidad en la que se analizan.

II. Objetivos

Objetivo general

Evaluar el uso de la Minería de Datos como una herramienta que sirva para la toma de decisiones a nivel gerencial.

Objetivos específicos

- Determinar en que consiste la minería de datos.
- Identificar los métodos y técnicas de la minería de datos.
- Conocer las herramientas principales de la minería de datos y sus propuestas para las organizaciones.
- Establecer si el proceso de la minería de datos permite generar conocimiento de la información analizada, para la toma de decisiones

III. Contenido

1. Conceptos de minería de datos, definiciones, características y objetivos.

1.1 Concepto de minería de datos

El data mining (minería de datos), es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Básicamente, el data mining surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento.

Para ello es importante conocer la diferencia entre datos, información y conocimiento **(1)**:

Datos: Los datos son la mínima unidad semántica, y se corresponden con elementos primarios de información que por sí solos son irrelevantes como apoyo a la toma de decisiones. También se pueden ver como un conjunto discreto de valores, que no dicen nada sobre el por qué de las cosas y no son orientativos para la acción.

Información: La información se puede definir como un conjunto de datos procesados y que tienen un significado, y que por lo tanto son de utilidad para quién debe tomar decisiones, al disminuir su incertidumbre. Los datos se pueden transforman en información añadiéndoles valor:

- Contextualizando: se sabe en qué contexto y para qué propósito se generaron.
- Categorizando: se conocen las unidades de medida que ayudan a interpretarlos.
- Calculando: los datos pueden haber sido procesados matemática o estadísticamente.
- Corrigiendo: se han eliminado errores e inconsistencias de los datos.
- Condensando: los datos se han podido resumir de forma más concisa (agregación).

Por tanto, la información es la comunicación de conocimientos o inteligencia, y es capaz de cambiar la forma en que el receptor percibe algo, impactando sobre sus juicios de valor y sus comportamientos.

Información = Datos + Contexto (añadir valor) + Utilidad (disminuir la incertidumbre)

***Conocimiento*:** El conocimiento es una mezcla de experiencia, valores, información y *know-how* que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción. Se origina y aplica en la mente de los conocedores. En las organizaciones con frecuencia no sólo se encuentra dentro de documentos o almacenes de datos, sino que también está en rutinas organizativas, procesos, prácticas, y normas.

El conocimiento se deriva de la información, así como la información se deriva de los datos. Para que la información se convierta en conocimiento es necesario realizar acciones como:

- Comparación con otros elementos.
- Predicción de consecuencias.
- Búsqueda de conexiones.
- Conversación con otros portadores de conocimiento.

1.2 Definiciones

El término “data mining” ha sido ampliamente utilizado por las empresas informáticas para identificar a productos y aplicaciones que, de forma genérica, analizan grandes cantidades de datos, con la finalidad de encontrar patrones o principios entre ellos. Se trataría de un amplio conjunto de técnicas utilizadas mediante una aproximación informática, cuya finalidad sería explorar y descubrir relaciones complejas en grandes conjuntos de datos **(2)**.

SAS Institute define el concepto de Data Mining como el proceso de Seleccionar (Selecting), Explorar (Exploring), Modificar (Modifying), Modelizar (Modeling) y Valorar (Assessment) grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores. Este proceso es resumido con las siglas SEMMA.

1.3 Características y objetivos

La minería de datos se caracteriza por[3]:

- Explorar los datos se encuentra en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años
- En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos; en otros, se mantienen en servidores de Internet e Intranet.
- El entorno de la minería de datos suele tener una arquitectura cliente servidor.
- Las herramientas de la minería de datos ayudan a extraer el mineral de la información enterrado en archivos corporativos o en registros públicos, archivados.
- Las herramientas de la minería de datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.
- Debido a la gran cantidad de datos, algunas veces resulta necesario usar procesamiento en paralelo para la minería de datos.

2. ¿Cómo Trabaja la minería de datos?[4]

¿Cómo es exactamente que la minería de datos es capaz, de extraer información importante que no se sabe que va a suceder después? La técnica que se utiliza para ejecutar estos hechos en la minería de datos se llama modelado de datos. El modelado datos es simplemente el acto de construir un modelo en una situación donde se sabe la respuesta y luego ésta se aplica a otra situación en la que no se sabe la respuesta. Por ejemplo, si se está buscando un tesoro de un galeón español en el mar lo primero que se debe hacer es investigar en el pasado cuando otros barcos encontraron tesoros. Se puede notar que estos buques a menudo tienden a encontrarse de las costas de Islas Bermudas y existen ciertas características de las corrientes de océano y ciertas rutas que probablemente fueron tomadas por los capitanes de los barcos en esa época. Si se toman estas similitudes y se construye un modelo que incluye las características que son comunes a las localizaciones de estos tesoros sumergidos, existe una gran posibilidad de que con estos modelos se pueda navegar en busca de otro tesoro, el modelo construido indica el lugar más probable donde pueda darse una situación similar al pasado. Si se ha construido un modelo adecuado, la probabilidad de encontrar un tesoro es bastante grande.

El proceso de construcción de un modelo es algo que las personas han estado haciendo durante mucho tiempo, desde antes del advenimiento de las computadoras. Lo que pasa con las computadoras, no es muy diferente de la forma en que las personas construían sus modelos. Las computadoras son cargadas con una gran cantidad de información sobre una variedad de situaciones donde se cuenta con una respuesta conocida y entonces el software de minería de datos en la computadora debe analizar a través de esos datos y determinar las características que se deben examinar por medio del modelo.

Una vez construido el modelo, éste puede ser usado en situaciones similares donde no se cuenta con una respuesta. Por ejemplo, si se supone que el director de marketing para una compañía de telecomunicaciones desea adquirir nuevos clientes que utilizan el teléfono para llamadas de larga distancia. De forma aleatoria se pueden enviar por correo cupones a la población general pero en ningún caso se logrará los resultados que se desean, ya que no toda la población realiza llamadas de larga distancia pero se puede utilizar la experiencia almacenada en la base de datos para construir un modelo.

Como el director de comercialización tiene acceso a una gran cantidad de información sobre todos sus clientes: edad, sexo, historia crediticia y uso de llamadas a larga distancia, se debería concentrar en aquellos usuarios que tienen un uso continuo de llamadas de larga distancia. Esto se puede realizar construyendo un modelo.

La meta al explorar los datos en busca de respuestas es hacer cierto cálculo sobre la información de los datos conocidos, el modelo que se construye va de la información general del cliente hacia información particular. Por ejemplo, un modelo simple para una compañía de telecomunicaciones podría ser El 98% de mis clientes que tiene un ingreso anual de más de S./.60,000.00 y gasta más de S/. 80.00 al mes en llamadas de larga distancia.

Este modelo se podría aplicar entonces a los datos para tratar de decir algo sobre la información con que se cuenta en la compañía de telecomunicaciones a la que normalmente no se tiene acceso. Con este modelo nuevos clientes pueden ser selectivamente fichados.

El data warehouse es una excelente fuente de datos para este tipo de modelado. Mirar los resultados de un data warehouse de prueba que representa una muestra grande pero relativamente pequeña de elementos que puede proporcionar fundamentos para identificar nuevos patrones en el data warehouse completo.

3. Aplicaciones de minería de datos[4]

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 están usando minería de datos para ayudar a manejar todas las fases del ciclo vital del cliente, incluyendo la adquisición de nuevos clientes, aumentando los ingresos con clientes existentes y manteniendo bien a la clientela.

Determinando características de clientes buenos, una compañía puede determinar conjuntos con características similares. Perfilando clientes que hayan comprado un producto en particular, ello puede enfocar la atención en clientes similares que no hayan comprado ese producto. 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 están haciendo mayor uso de la minería de datos, para decidir que productos en particular deben mantener en inventario para no abastecerse de productos innecesarios, así como para evaluar la eficacia de promociones y ofertas.

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

4. Metodología CRISP-DM[5]

CRISP-DM, está dividida en 4 niveles de abstracción organizados de forma jerárquica (figura 1) en tareas que van desde el nivel más general, hasta los casos más específicos y organiza el desarrollo de un proyecto de Data Mining (DM), en una serie de seis fases (figura 2):

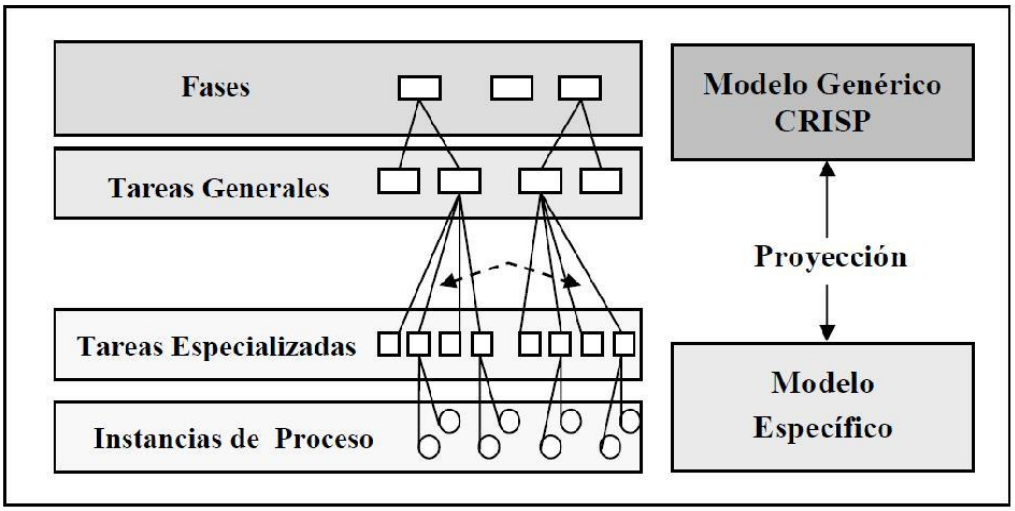

Figura 1: Esquema de los 4 niveles de CRISP-DM[6].

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.

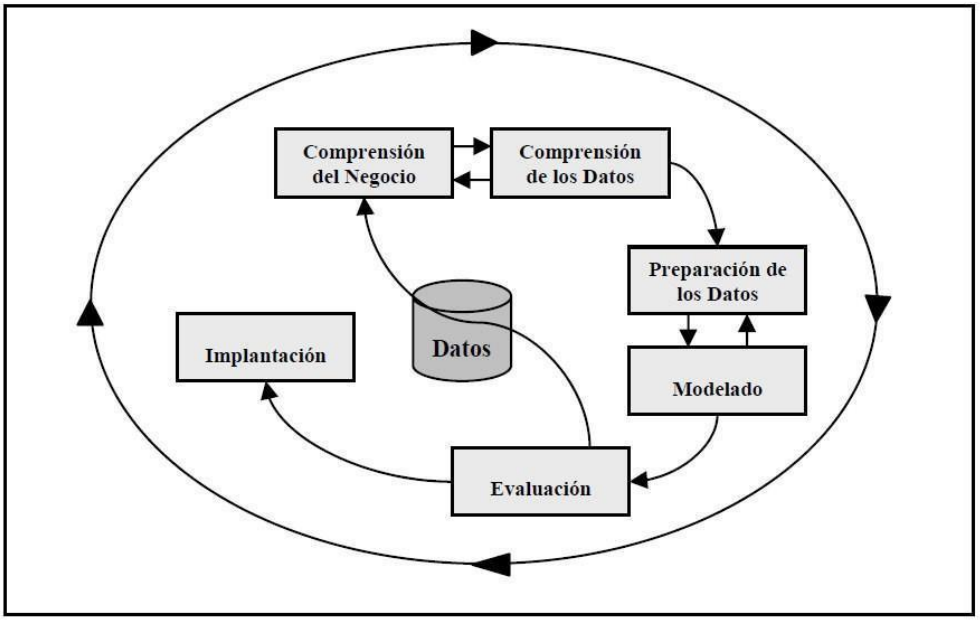

Figura 2: Modelo de proceso CRISP-DM[6].

A continuación se describen cada una de las fases en que se divide CRISP-DM.

4.1 Fase de compresión del negocio o problema

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema (figura 3), es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de Data Mining, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Data Mining y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. Una descripción de cada una de las principales tareas que componen esta fase es la siguiente:

Determinar los objetivos del negocio. Esta es la primera tarea a desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Data Mining y definir los criterios de éxito. Los problemas pueden ser diversos como por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio, califica el resultado del proceso de DM, o de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.

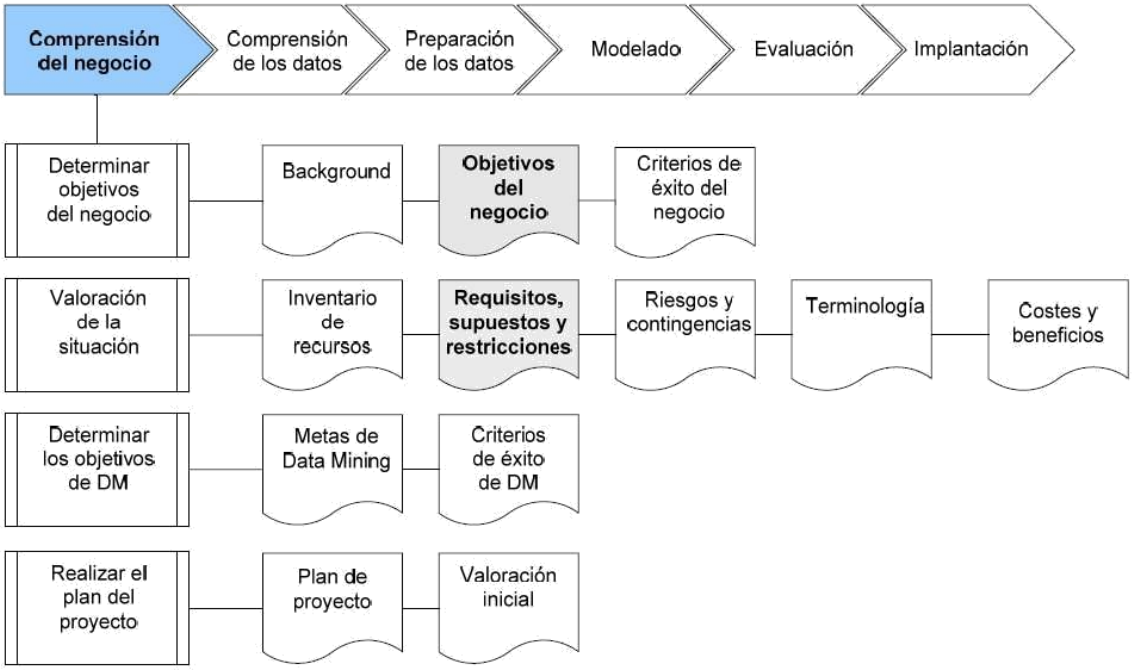

Figura 3: Fase de compresión del negocio [6].

Evaluación de la situación. En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de DM, considerando aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de DM?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de Data Mining.

Determinación de los objetivos de DM. Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de DM, como por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de DM será por ejemplo, determinar el perfil de los clientes respecto de su capacidad de endeudamiento. Producción de un plan del proyecto. Finalmente esta última tarea de la primera fase de CRISP-DM, tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada paso.

4.2 Fase de compresión de datos

La segunda fase (figura 4), fase de comprensión de los datos, 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. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas.

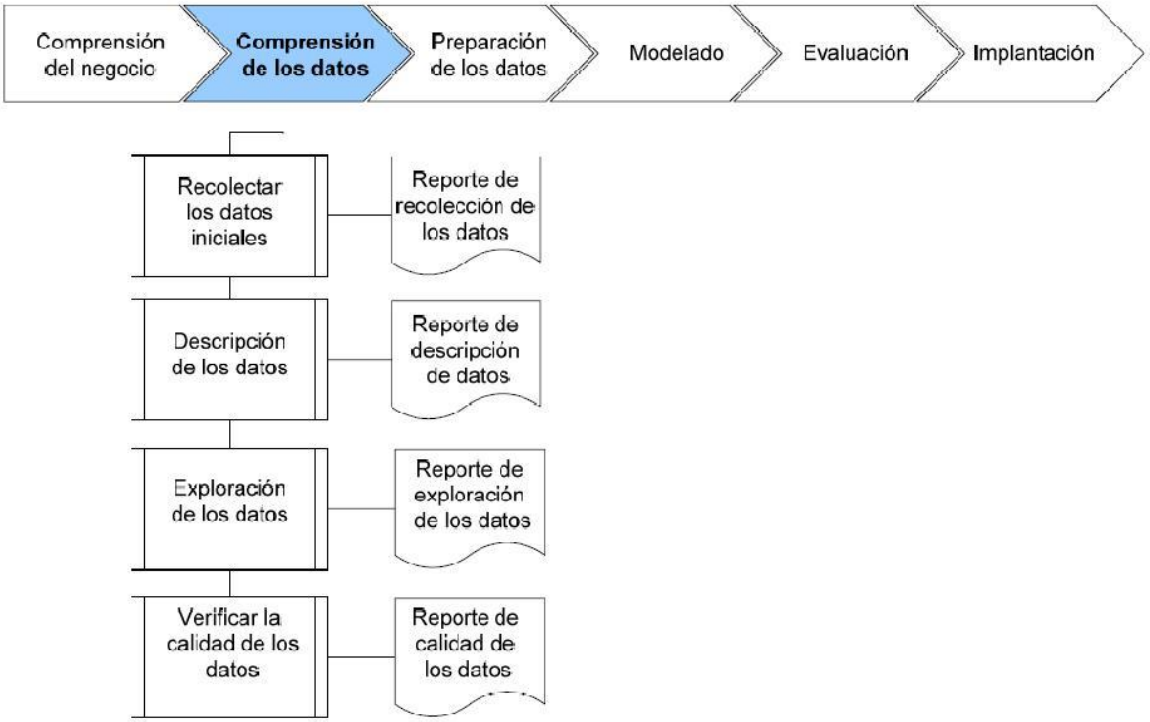

Figura 4: Fase de compresión de los datos [6].

Las principales tareas a desarrollar en esta fase del proceso son:

Recolección de datos iniciales. La primera tarea en esta segunda fase del proceso de CRISP-DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.

Descripción de los datos. Después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

Exploración de datos. A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.

Verificación de la calidad de los datos. En esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea en este punto, es asegurar la completitud y corrección de los datos.

4.3 Fase de preparación de los datos

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos 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.

Esta fase se encuentra relacionada con la fase de modelado, puesto que en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas. Es así que las fases de preparación y modelado interactúan de forma permanente. La figura 5, ilustra las áreas de que se compone ésta, e identifica sus salidas. Una descripción de las tareas involucradas en esta fase es la siguiente:

Selección de datos. En esta etapa, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas.

Limpieza de los datos. Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.

Estructuración de los datos. Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.

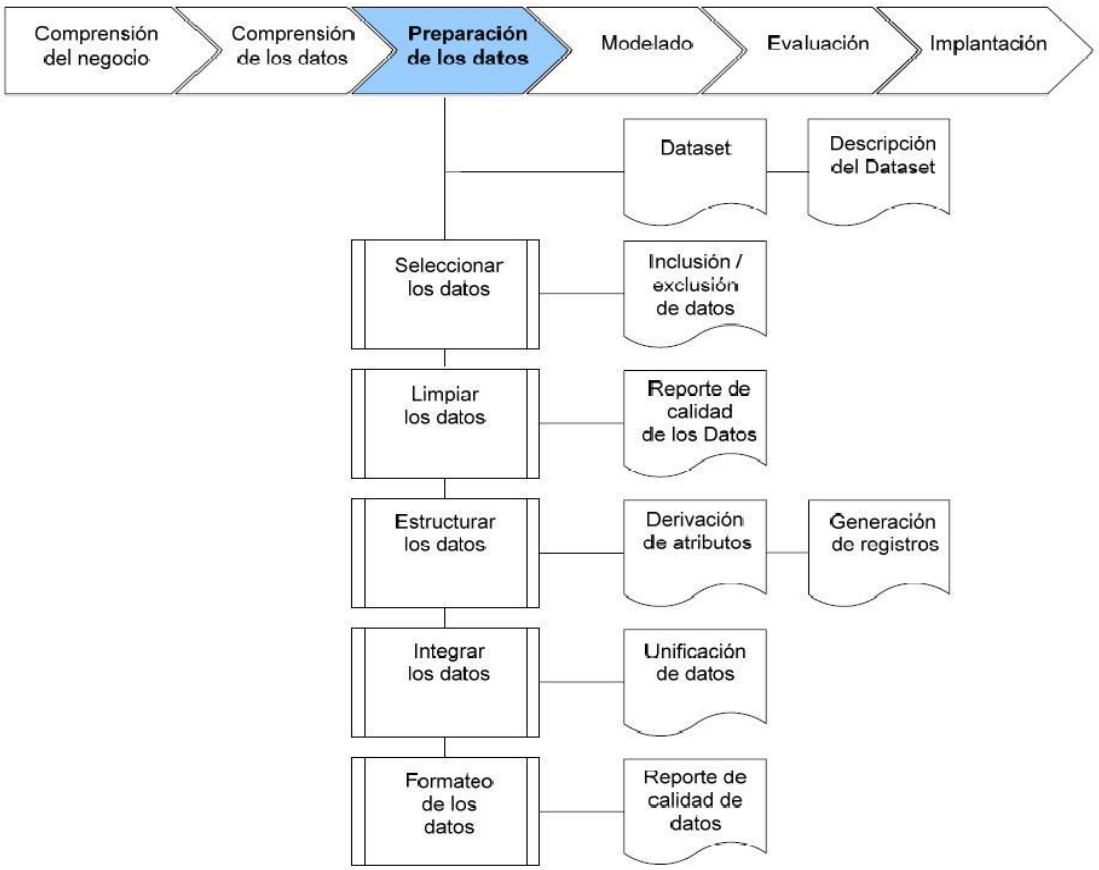

Figura 5: Fase de preparación de los datos[6].

Integración de los datos. La integración de los datos, involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

Formateo de los datos. Esta tarea consiste principalmente, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de DM en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

4.4 Fase de modelado

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase 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.

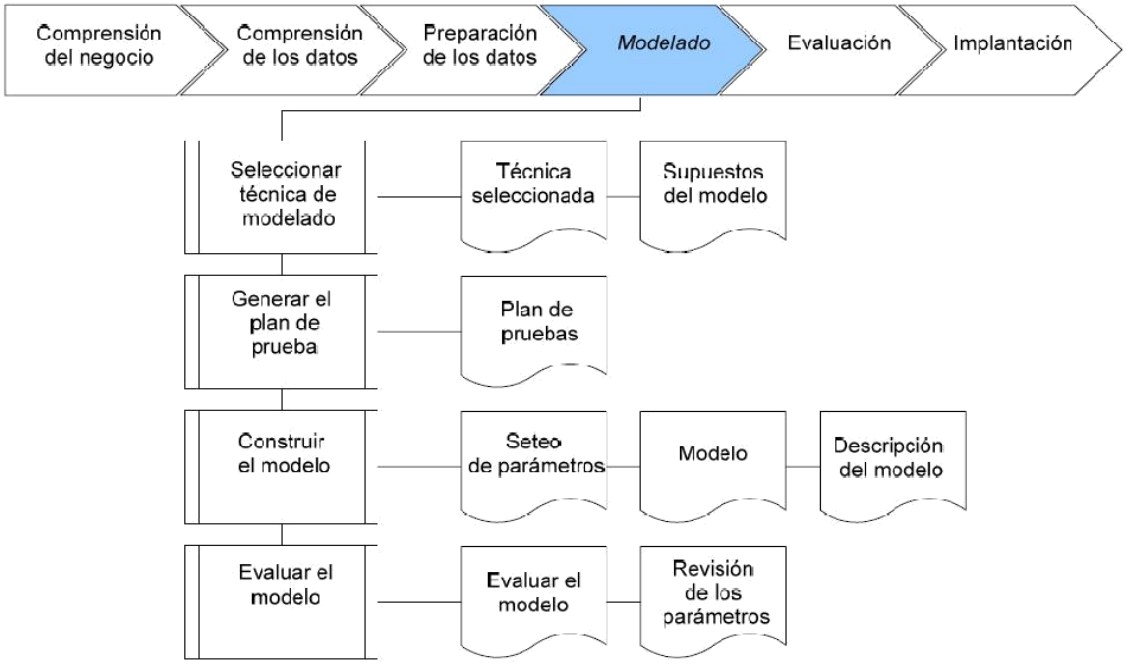

Figura 6: Fase de modelado [6].

Previamente al modelado de los datos, se debe determinar un método de evaluación de los modelos que permita establecer el grado de bondad de ellos. Después de concluir estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo, dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo. La figura 7 ilustra las tareas y resultados que se obtienen en esta fase. Una descripción de las principales tareas de esta fase es la siguiente:

Selección de la técnica de modelado. Esta tarea consiste en la selección de la técnica de DM más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de DM existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, razonamiento basado en casos (CBR); si el problema es de predicción, análisis de regresión, redes neuronales; o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.

Generación del plan de prueba. Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Por ejemplo, en una tarea supervisada de DM como la clasificación, es común usar la razón

de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

Construcción del Modelo. Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

Evaluación del modelo. En esta tarea, los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Data Mining aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc...).

4.5 Fase de evaluación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis.

Considerar que se pueden emplear múltiples herramientas para la implementación de los resultados son muy empleadas en problemas de clasificación y consisten en una tabla que indica cuantas clasificaciones se han hecho para cada tipo, la diagonal de la tabla representa las clasificaciones correctas. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo. La figura 7 detalla las tareas que componen esta fase y los resultados que se deben obtener. Las tareas involucradas en esta fase del proceso son las siguientes:

Evaluación de los resultados. En los pasos de evaluación anteriores, se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente,

o si es aconsejable probar el modelo, en un problema real si el tiempo y restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.

Proceso de revisión. El proceso de revisión, se refiere a calificar al proceso entero de DM, a objeto de identificar elementos que pudieran ser mejorados.

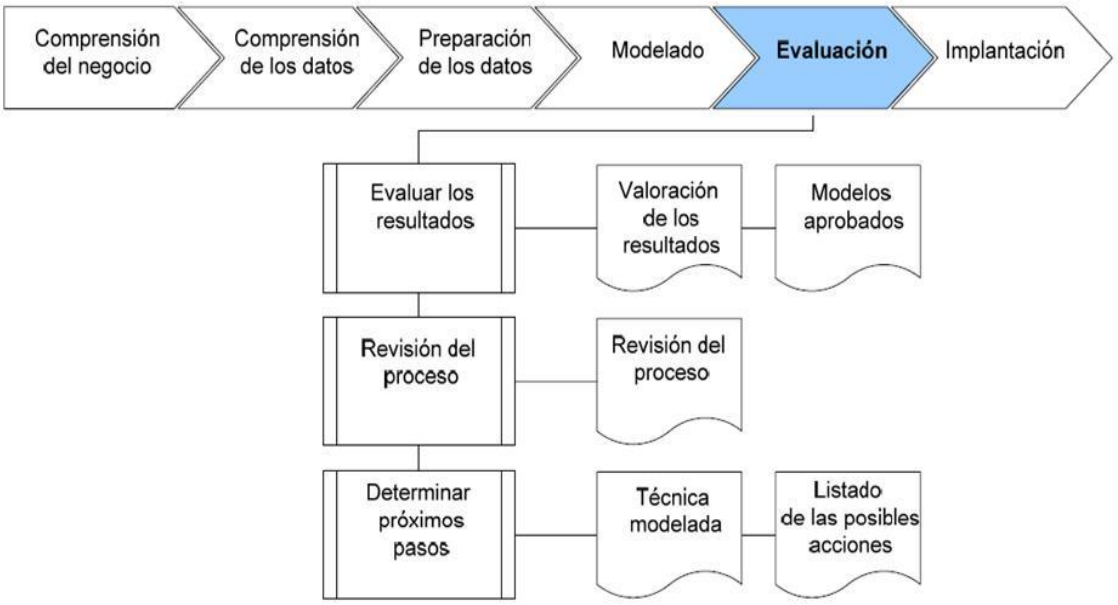

Figura 7: Fase de evaluación[6].

Determinación de futuras fases. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de DM.

4.6 Fase de implementación

En esta fase (figura 8), 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, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, como por ejemplo, en análisis de riesgo crediticio, detección de fraudes, etc. Generalmente un proyecto de Data Mining

no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Las tareas que se ejecutan en esta fase son las siguientes:

Plan de implementación. Para implementar el resultado de DM en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación. Monitorización y Mantenimiento. Si los modelos resultantes del proceso de Data Mining son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

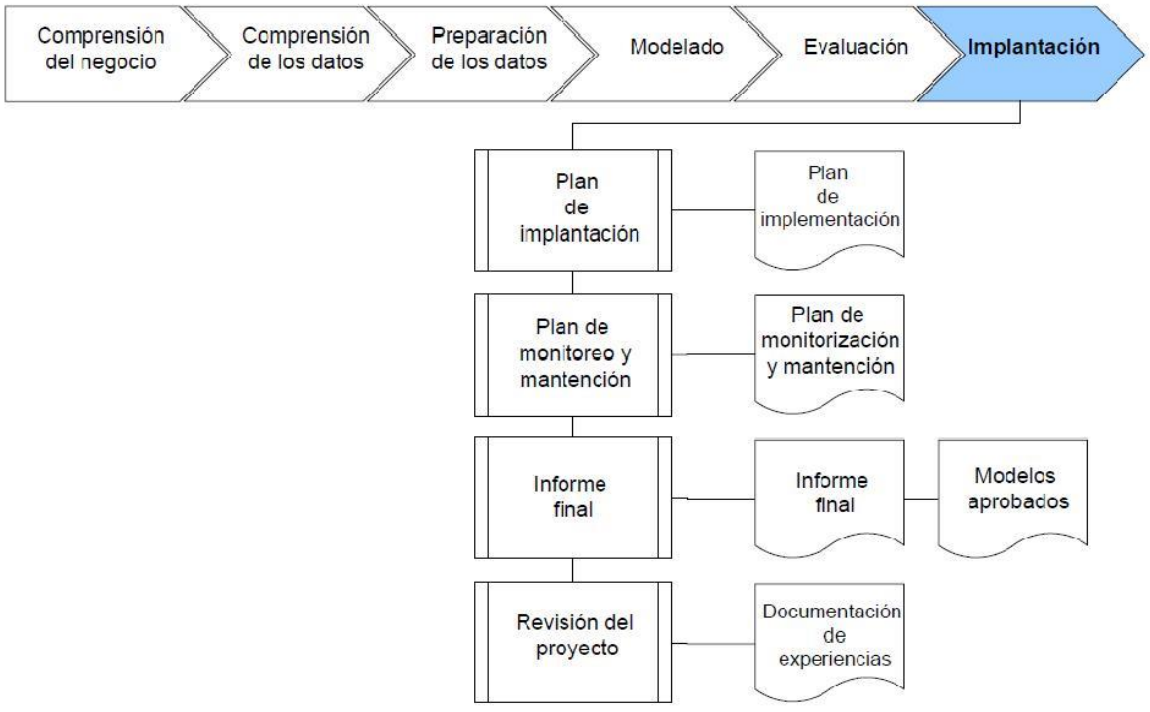

Figura 8: Fase de implementación [6].

Informe Final. Es la conclusión del proyecto de DM realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto. Revisión del proyecto: En este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar.

5. Otros procesos y metodologías estandarizados de minería de datos.

CRISP–DM [5], es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining, como se puede constatar en la gráfica presentada en la figura 9. Esta gráfica, publicada el año 2007 por kdnuggets.com, representa el resultado obtenido en sucesivas encuestas efectuadas durante los últimos años, respecto del grado de utilización de las principales guías de desarrollo de proyectos de Data Mining. En ella se puede observar, que a pesar de que el uso de CRISP-DM ha descendido desde un 51 % en el año 2002, a un 42% en el año 2007, es aun frente a otras, la guía de referencia más ampliamente utilizada.

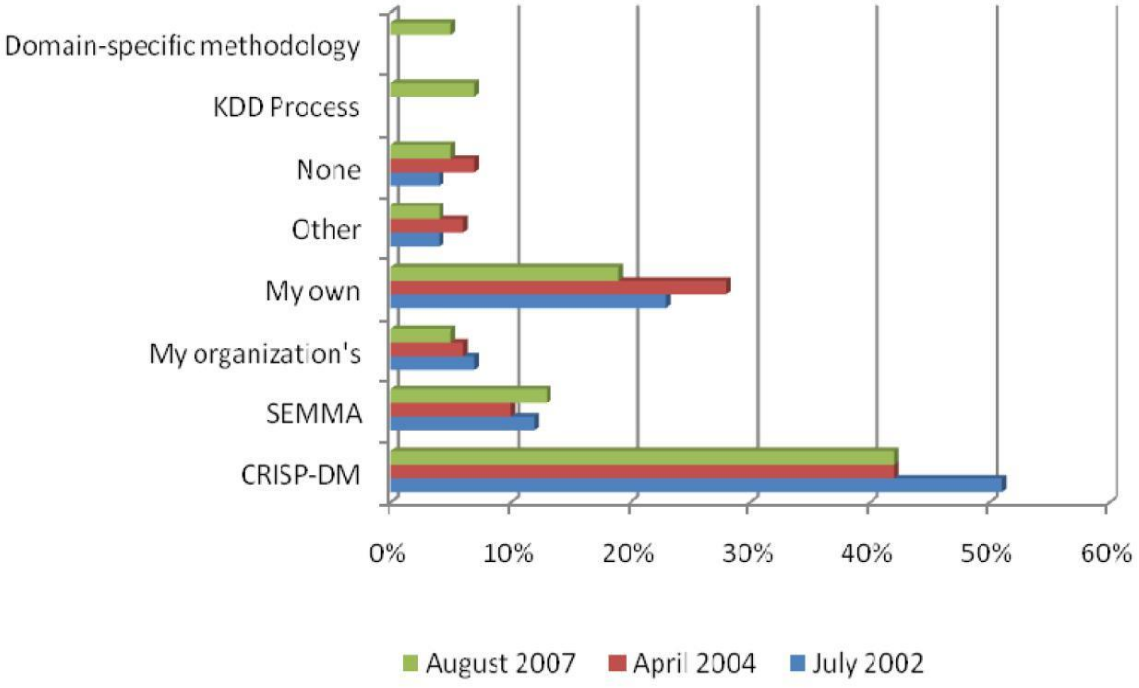

Figura 9: Metodologías utilizadas en Data Mining [7].

5.1 SEMMA[8]

SAS Institute, es el desarrollador de esta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos.

El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso (Figura 10).

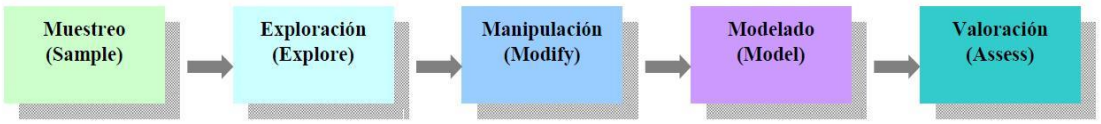

Figura 10: Fases de la metodología SEMMA

El proceso se inicia con la 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.

La metodología SEMMA establece que para cada muestra considerada para el análisis del proceso se debe asociar el nivel de confianza de la muestra. Una vez determinada una muestra o conjunto de muestras representativas de la población en estudio, la metodología SEMMA indica que se debe proceder a una exploración de la información disponible con el fin de simplificar en lo posible el problema para 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.

La tercera fase de la metodología 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. Las técnicas utilizadas 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 (difusa), árboles de decisión, reglas de asociación y computación evolutiva.

Finalmente, la última fase del proceso consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos contrastados con otros métodos estadísticos o con nuevas poblaciones muestrales.

5.2 Microsoft[9]

En la Figura 11 se describe las relaciones entre cada paso en la metodología desarrollada por Microsoft para la implementación de Data Mining (Figura 11).

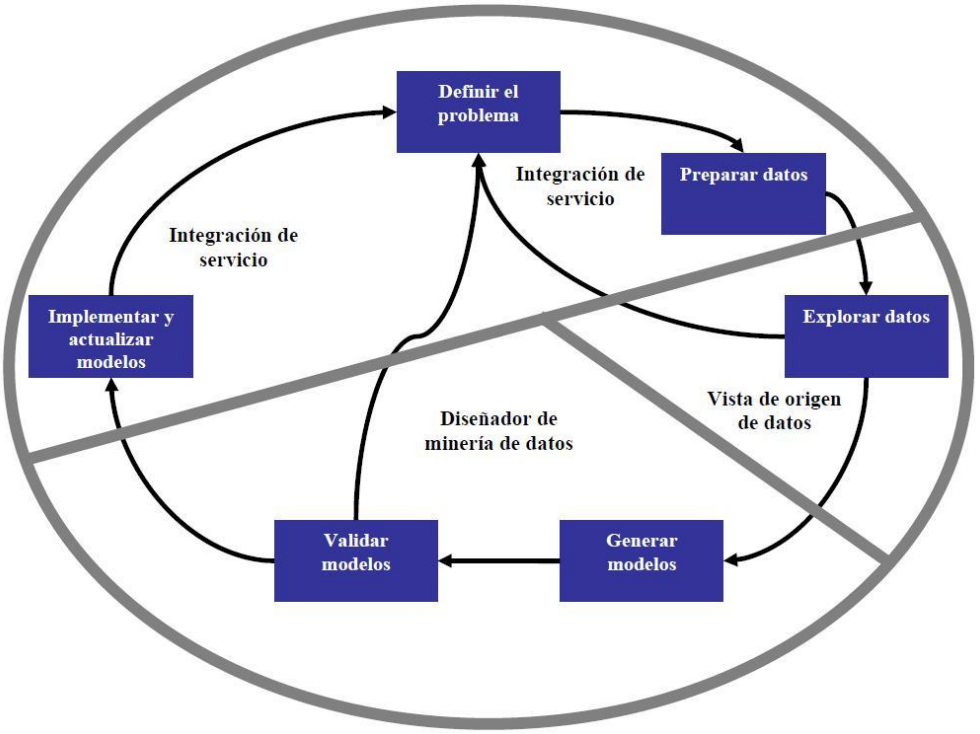

Figura 11: Fases del proceso de modelado metodología Microsoft

El *primer paso* del proceso de minería de datos consiste en definir claramente el problema empresarial. Este paso incluye analizar los requisitos empresariales, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir el objetivo final del proyecto de minería de datos. Estas tareas se traducen en preguntas como las siguientes:

- ¿Qué se está buscando?
- ¿Qué atributo del conjunto de datos se desea intentar predecir?
- ¿Qué tipos de relaciones se intenta buscar?
- ¿Se desea realizar predicciones a partir del modelo de minería de datos o sólo buscar asociaciones y patrones interesantes?
- ¿Cómo se distribuyen los datos?
- ¿Cómo se relacionan las columnas? o en caso de haber varias tablas, ¿cómo se relacionan las tablas?

Para responder a estas preguntas, es probable que se deba dirigir un estudio de disponibilidad de datos para investigar las necesidades de los usuarios de la empresa con respecto a los datos disponibles. Si los datos no son compatibles con las necesidades de los usuarios, puede que se deba volver a definir el proyecto.

El *segundo paso* del proceso de minería de datos consiste en consolidar y limpiar los datos identificados en el paso “Definir el problema”.

Los datos pueden estar dispersos en la empresa y almacenados en distintos formatos; también pueden contener incoherencias como entradas que faltan o contienen errores.

El *tercer paso* del proceso de minería de datos consiste en explorar los datos preparados. Se debe comprender los datos para tomar las decisiones adecuadas al crear los modelos. Entre las técnicas de exploración se incluyen calcular los valores mínimos y máximos, calcular la media y las desviaciones estándar y examinar la distribución de los datos. Una vez explorados los datos, se puede decidir si el conjunto de datos contiene datos con errores y, a continuación, crear una estrategia para solucionar los problemas.

El *cuarto paso* del proceso de minería de datos consiste en generar los modelos de minería de datos. Antes de generar un modelo, se deben separar aleatoriamente los datos preparados en conjuntos de datos de entrenamiento y comprobación independientes. El conjunto de datos de entrenamiento se utiliza para generar el modelo y el conjunto de datos de comprobación para comprobar la precisión del modelo mediante la creación de consultas de predicción.

Una vez definida la estructura del modelo de minería de datos, se procesa rellenando la estructura vacía con los patrones que describen el modelo. Esto se conoce como entrenar el modelo. Los patrones se encuentran al pasar los datos originales por un algoritmo matemático.

El modelo de minería de datos se define mediante un objeto de estructura de minería de datos, un objeto de modelo de minería de datos y un algoritmo de minería de datos.

El *quinto paso* del proceso de minería de datos consiste en explorar los modelos que se han generado y comprobar su eficacia. No se debe implementar un modelo en un entorno de producción sin comprobar primero si el modelo funciona correctamente. Además, puede ser que se hayan creado varios modelos y se deba decidir cuál funciona mejor. Si ninguno de los modelos que se han creado en el paso Generar Modelos funciona correctamente, puede ser que se deba volver a un paso anterior del proceso y volver a definir el problema o volver a investigar los datos del conjunto de datos original.

El *último paso* del proceso de minería de datos consiste en implementar los modelos que funcionan mejor en un entorno de producción.

Una vez que los modelos de minería de datos se encuentran en el entorno de producción, se pueden llevar acabo diferentes tareas, dependiendo de las necesidades.

Éstas son algunas de las tareas que se pueden realizar:

- Utilizar los modelos para crear predicciones que se puedan utilizar para tomar decisiones empresariales.
- Incrustar la funcionalidad de minería de datos directamente en una aplicación.

- Crear un paquete en el que se utilice un modelo de minería de datos para separar de forma inteligente los datos entrantes en varias tablas.
- Crear un informe que permita a los usuarios realizar consultas directamente en un modelo de minería de datos existente.

La actualización del modelo forma parte de la estrategia de implementación. A medida que la organización recibe más datos, se deben volver a procesar los modelos para mejorar así su eficacia.

5.3 Comparación de metodologías

Las metodologías SEMMA, CRISP-DM y Microsoft esencialmente son muy parecidas. Las tres están compuestas por etapas o fases que interactúan entre sí.

En referencia a la tecnología SEMMA está más ligada a los aspectos técnicos de la explotación de datos. En cuanto a las otras dos, están más enfocadas 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 y Microsoft empiezan por el estudio del negocio y sus objetivos, y luego recién se transforma en un problema técnico.

Analizando la propuesta metodológica de Microsoft se puede ver que está íntimamente vinculada a la aplicación de las herramientas de su propia compañía (Microsoft) especialmente en lo que respecta a la integración de servicios, vista de origen de datos y diseñador de minería de datos. Lo mismo ocurre con la metodología SEMMA la cual está ligada a herramientas SAS.

Para concluir se puede decir que uno de los motivos por los cuales fue escogida 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.

Finalmente también 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 1 muestra un cuadro comparativo entre las diferentes metodologías descriptas hasta aquí.

Metodologías	CRISP-DM	SEMMA	Microsoft
Estructura	Fases y niveles	Fases	Fases
Niveles	Parte de lo general a lo específico	No tiene	No tiene
Fases	Análisis del problema Análisis de datos Preparación de Datos Modelado Evaluación Explotación	Muestreo Exploración Manipulación Modelado Valoración	Definir el problema Preparar los datos Explorar los datos Generar modelos Explorar y validar los modelos Implementar y actualizar los modelos
Herramientas	Genéricas	SAS	Microsoft
Procesos	Iterativo e interactivo entre fases	Iterativo e interactivo entre fases	Iterativo e interactivo entre fases
Documentación	Modelo de referencia Guía de usuario	No se especifica	No se especifica
Objetivos	Se centra en los objetivos empresariales del proyecto	Se centra en las características técnicas del desarrollo del proceso	Se centra en los objetivos empresariales del proyecto

Tabla 1: Cuadro comparativo de metodologías

6. Modelos y tareas de minería de datos

Al ser la minería de datos un método para extraer conocimiento útil mediante el análisis de los datos, ésta recurre a modelos que permitan encontrar relaciones, patrones o reglas inferidas previamente desconocidas **(11).** Los modelos empleados en la minería de datos son el descriptivo y el predictivo.

6.1 Modelo descriptivo

En el modelo descriptivo se identifican patrones que describen los datos mediante tareas, ej. Agrupamiento (clustering) y reglas de asociación **(12). (11)** destacan que mediante este modelo se identifican patrones que explican o resumen el conjunto de datos, siendo estos útiles para explorar las propiedades de los datos examinados. Los modelos descriptivos siguen un tipo de aprendizaje no supervisado, que consiste en adquirir conocimiento desde los datos disponibles, sin requerir influencia externa que indique un comportamiento

deseado al sistema

6.2 Modelo predictivo

Este modelo se emplea para estimar valores futuros de variables de interés. El proceso se basa en la información histórica de los datos, mediante las cuales se predice un comportamiento de los datos, ya sea mediante clasificaciones, categorizaciones o regresiones **(11,13).** Los modelos predictivos siguen un aprendizaje supervisado, que consiste en aprender mediante el control de un supervisor o maestro que determina la respuesta que se desea generar del sistema **(13)**. El atributo a predecir se conoce como variable dependiente u objetivo, mientras que los atributos utilizados para realizar la predicción se llaman variables independientes o de exploración **(11).**

6.3 Tareas de minería de datos

Dentro de los modelos descriptivos y predictivos se encuentran diferentes tareas específicas como: agrupamiento, reglas de asociación, clasificación, regresión, entre otras. Estas tareas corresponden a un tipo de problema específico en el proceso de minería de datos. En la Figura 11, se muestra una representación general de los modelos y tareas hallados en el proceso de minería de datos, siendo abordadas aquí brevemente cada una de ellas.

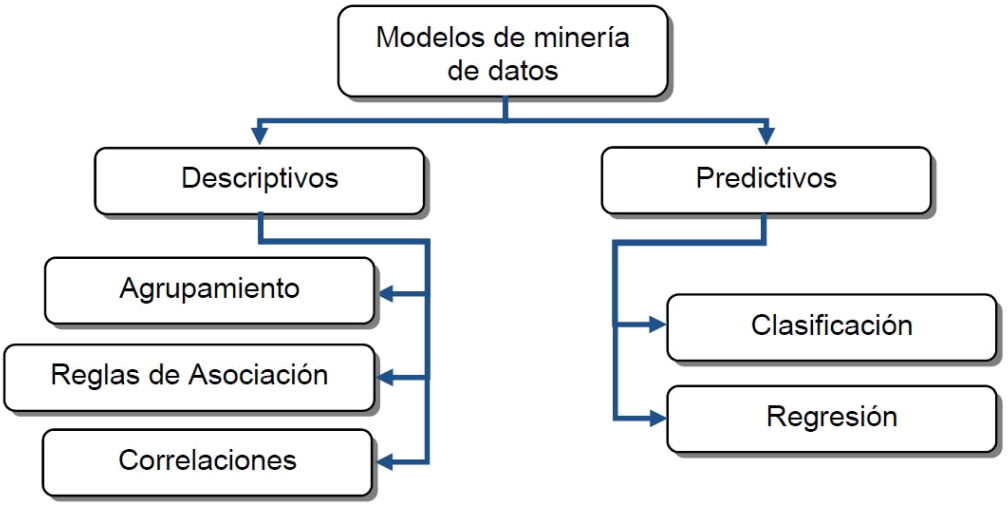

Figura 12: Representación general de los modelos y tareas de minería de datos[10]

6.3.1 Agrupamiento (Clustering)

En esta tarea se evalúan similitudes entre los datos para construir modelos descriptivos, analizar correlaciones entre las variables o representar un conjunto de datos en un pequeño número de regiones **(13).**Berry y Linoff y Sumathi y Sivanandam consideran al agrupamiento como la tarea de dividir una población heterogénea en un número de subgrupos homogéneos de acuerdo a las similitudes de sus registros **(13,14)**. Dentro de esta tarea existen dos tipos principales de agrupamiento **(15)**: el *jerárquico* que se caracteriza por el desarrollo recursivo de una estructura en forma de árbol, y el *particional* que organiza los registros dentro de *k* grupos. Los métodos particionales tienen ventajas en aplicaciones que involucran gran cantidad de datos para los cuales la construcción de un árbol resulta complicada. Una característica de este tipo de agrupamiento es el establecer a priori el número de grupos de entrada (*k*), por lo que en la práctica es necesario repetir la prueba estableciendo diferentes números de grupos, eligiendo la solución que mejor se adapte al objetivo del problema **(16)**. Un método sugerido por Milligan (1980, 1985) y Hair et al. (1995) para determinar el número de grupos de entrada (*k*) es usar el resultado obtenido por algún algoritmo jerárquico, mediante el cual se obtiene el número deseado de grupos, posteriormente se aplica algún algoritmo particional.

6.3.2 Reglas de asociación

Mediante esta tarea se identifican afinidades entre la colección de los registros examinados, buscando relaciones o asociaciones entre ellos. Las afinidades son expresadas como reglas de la forma: “Si X entonces Y”, donde X y Y son los registros de una transacción **(13)**. El interés por esta tarea se debe principalmente a que las reglas proporcionan una forma concisa de declarar la información potencialmente útil **(12)**. Las reglas se evalúan usando dos parámetros: precisión y cobertura. La *cobertura* es el número de instancias o datos hallados correctamente, mientras que la *precisión* es el porcentaje de instancias halladas correctamente **(17)**. Las ventajas más frecuentes en las reglas de asociación son el *descubrimiento de asociación y de secuencia* **(11)**. El descubrimiento de asociación encuentra relaciones que aparecen conjuntamente a un acontecimiento y la secuencia la asocia al tiempo.

6.3.3 Correlaciones

Las correlaciones son una tarea descriptiva que se usan para determinar el grado de similitud de los valores de dos variables numéricas. Un mecanismo estándar para medir la correlación es el coeficiente de correlación, para este caso llamado *R*, el cual es un valor real comprendido entre -1 y 1. Así, si *R* es 1 las variables están totalmente correlacionadas; si R es -1 las variables están correlacionadas negativamente; y si *R* es cero no existe correlación. Por consiguiente, cuando *R* es positivo, las variables tienen un comportamiento similar y cuando *R* es negativo una variable crece y la otra decrece **(11)**. Una manera de visualizar la posible correlación entre las observaciones de dos variables (X e Y), es a través de un diagrama de dispersión, en el cual los valores que toman estas variables son representados por puntos. Su principal desventaja es que no puede ser usada para hacer predicciones, puesto que no es clara la forma que toma la relación.

6.3.4 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 **(12)**. 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 **(13)**. 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 **(12, 18)**.

6.3.5 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 **(13)**. En la regresión, la información de salida es un valor numérico continuo o un vector con valores no discretos **(12)**. 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, cuando el valor o valores predecidos están en medio de otros; o estimación, cuando se predice valores futuros **(11)**.

7. Redes Neuronales Artificiales[18]

Las Redes Neuronales Artificiales (RNA) o sistemas conexionistas son sistemas de procesamiento de la información cuya estructura y funcionamiento están inspirados en las redes neuronales biológicas. Consisten en un conjunto de elementos simples de procesamiento llamados nodos o neuronas conectadas entre sí por conexiones que tienen un valor numérico modificable llamado peso.

La actividad que una unidad de procesamiento o neurona artificial realiza en un sistema de este tipo es simple. Normalmente, consiste en sumar los valores de las entradas (inputs) que recibe de otras unidades conectadas a ella, comparar esta cantidad con el valor umbral y, si lo iguala o supera, enviar activación o salida (output) a las unidades a las que esté conectada. Tanto las entradas que la unidad recibe como las salidas que envía dependen a su vez del peso o fuerza de las conexiones por las cuales se realizan dichas operaciones.

La arquitectura de procesamiento de la información de los sistemas de RNA se distingue de la arquitectura convencional Von Neumann (fundamento de la mayor parte de los ordenadores existentes) en una serie de aspectos fundamentales:

En primer lugar, el procesamiento de la información de un modelo Von Neumann es secuencial, esto es, una unidad o procesador central se encarga de realizar una tras otra determinadas transformaciones de expresiones binarias almacenadas en la memoria del ordenador. Estas transformaciones son realizadas de acuerdo con una serie de instrucciones (algoritmo, programa), también almacenadas en la memoria. La operación básica de un sistema de este tipo sería: localización de una expresión en la memoria, traslado de dicha expresión a la unidad de procesamiento, transformación de la expresión y colocación de la nueva expresión en otro compartimento de la memoria. Por su parte, el procesamiento en un sistema conexionista no es secuencial sino paralelo, esto es, muchas unidades de procesamiento pueden estar funcionando simultáneamente.

En segundo lugar, un rasgo fundamental de una arquitectura Von Neumann es el carácter discreto de su memoria, que está compuesta por un gran número de ubicaciones físicas o compartimentos independientes donde se almacenan en código digital tanto las instrucciones (operaciones a realizar) como los datos o números que el ordenador va a utilizar en sus operaciones. En redes neuronales, en cambio, la información que posee un sistema no está localizada o almacenada en compartimentos discretos, sino que está distribuida a lo largo de los parámetros del sistema. Los parámetros que definen el "conocimiento" que una red neuronal posee en un momento dado son sus conexiones y el estado de activación de sus unidades de procesamiento. En un sistema conexionista las expresiones lingüísticas o simbólicas no existen como tales. Serían el resultado emergente de la interacción de muchas unidades en un nivel sub simbólico.

Por último, un sistema de RNA no se programa para realizar una determinada tarea a diferencia de una arquitectura Von Neumann, sino que es "entrenado" a tal efecto. Consideremos un ejemplo típico de aprendizaje o formación de conceptos en la estructura de una RNA. Supongamos que presentamos a la red dos tipos de objetos, por ejemplo la letra A y la letra E con distintos tamaños y en distintas posiciones. En el aprendizaje de la red neuronal se consigue, tras un número elevado de presentaciones de los diferentes objetos y consiguiente ajuste o modificación de las conexiones del sistema, que la red distinga entre As y Es, sea cual fuere su tamaño y posición en la pantalla. Para ello, podríamos entrenar la red neuronal para que proporcionase como salida el valor 1 cada vez que se presente una A y el valor 0 en caso de que se presente una E. El aprendizaje en una RNA es un proceso de ajuste o modificación de los valores o

pesos de las conexiones, “hasta que la conducta del sistema acaba por reproducir las propiedades estadísticas de sus entradas” **(20)**. En nuestro ejemplo, podríamos decir que la red ha “aprendido” el concepto de letra A y letra
E sin poseer reglas concretas para el reconocimiento de dichas figuras, sin poseer un programa explícito de instrucciones para su reconocimiento.

Por tanto, para entrenar a un sistema conexionista en la realización de una determinada clasificación es necesario realizar dos operaciones. Primero, hay que seleccionar una muestra representativa con respecto a dicha clasificación, de pares de entradas y sus correspondientes salidas. Segundo, es necesario un algoritmo o regla para ajustar los valores modificables de las conexiones entre las unidades en un proceso iterativo de presentación de entradas, observación de salidas y modificación de las conexiones.

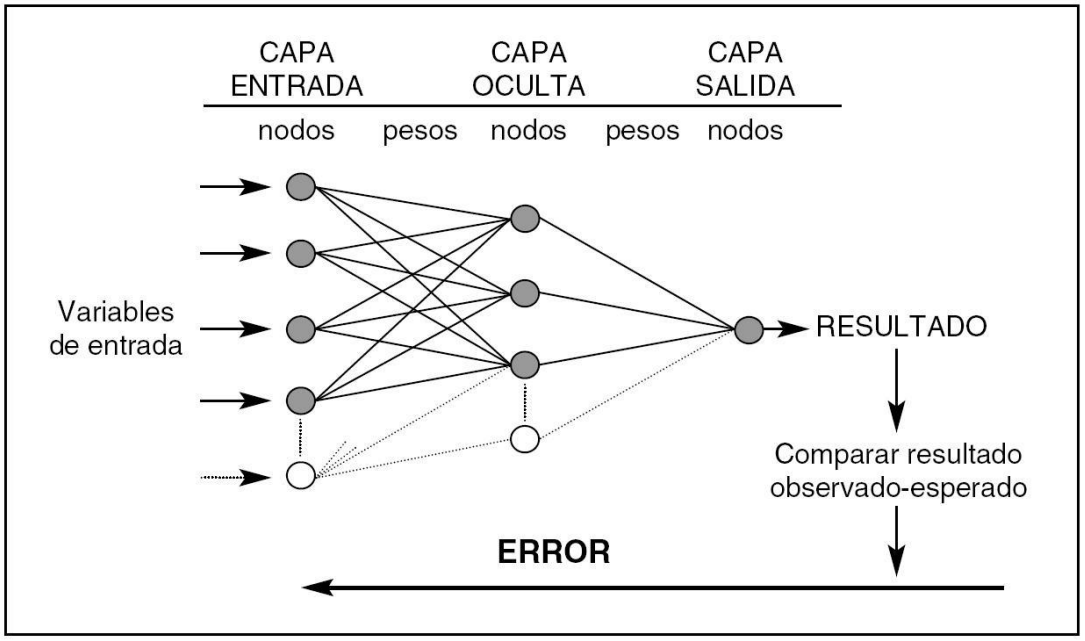

Figura 13: Arquitectura de una red neuronal artificial (RNA) **21**

7.1 Elementos de una RNA

Todas las RNA tienen unos elementos en común que son los siguientes[22]:

- **Neuronas** y los elementos que la forman: valor, señal de salida, peso de la sinapsis (factor asignado a cada sinapsis), entrada total, función de salida, función de activación y reglas de aprendizaje (permiten modificar los pesos de la sinapsis).
- **Capa o nivel**: conjunto de neuronas cuya capa tiene su origen en la misma fuente y cuyas salidas van al mismo destino.
- **Tipos de capas**: entrada (reciben estímulos externos), ocultas (representación interna de la información) y salida.
- **Conexión entre neuronas**: propagación hacia delante (ninguna salida de las neuronas es entrada del mismo nivel o niveles superiores) y propagación hacia detrás (la salida de las neuronas pueden ser entradas del mismo nivel o niveles anteriores y también de ellas mismas).
- **Dinámica**: asincrónica (evalúan su estado continuamente, según les llega información), sincronía (cambios a la vez en todas las neuronas).

7.2 Aplicaciones

Las redes neuronales pueden utilizarse en un gran número y variedad de aplicaciones, tanto comerciales como militares.

Se pueden desarrollar redes neuronales en un periodo de tiempo razonable, con la capacidad de realizar tareas concretas mejor que otras tecnologías. Cuando se implementan mediante hardware (redes neuronales en chips VLSI), presentan una alta tolerancia a fallos del sistema y proporcionan un alto grado de paralelismo en el procesamiento de datos. Esto posibilita la inserción de redes neuronales de bajo coste en sistemas existentes y recientemente desarrollados.

Hay muchos tipos diferentes de redes neuronales; cada uno de los cuales tiene una aplicación particular más apropiada. Algunas aplicaciones comerciales son[23]:

- Biología:
  - Aprender más acerca del cerebro y otros sistemas.
  - Obtención de modelos de la retina.

- Empresa:
  - Evaluación de probabilidad de formaciones geológicas y petrolíferas.
  - Identificación de candidatos para posiciones específicas.
  - Explotación de bases de datos.
  - Optimización de plazas y horarios en líneas de vuelo.
  - Optimización del flujo del tránsito controlando convenientemente la temporización de los semáforos.
  - Reconocimiento de caracteres escritos.
  - Modelado de sistemas para automatización y control.

- Medio ambiente:
  - Analizar tendencias y patrones.
  - Previsión del tiempo.

- Finanzas:
  - Previsión de la evolución de los precios.
  - Valoración del riesgo de los créditos.
  - Identificación de falsificaciones.
  - Interpretación de firmas.

- Manufacturación:
  - Robots automatizados y sistemas de control (visión artificial y sensores de presión, temperatura, gas, etc.).
  - Control de producción en líneas de procesos.
  - Inspección de la calidad.

- Medicina:
  - Analizadores del habla para ayudar en la audición de sordos profundos.
  - Diagnóstico y tratamiento a partir de síntomas y/o de datos analíticos (electrocardiograma, encefalogramas, análisis sanguíneo, etc.).

- Monitorización en cirugías.
- Predicción de reacciones adversas en los medicamentos.
- Entendimiento de la causa de los ataques cardíacos.

- Militares:
  - Clasificación de las señales de radar.
  - Creación de armas inteligentes.
  - Optimización del uso de recursos escasos.
  - Reconocimiento y seguimiento en el tiro al blanco.

La mayoría de estas aplicaciones consisten en realizar un reconocimiento de patrones, como ser: buscar un patrón en una serie de ejemplos, clasificar patrones, completar una señal a partir de valores parciales o reconstruir el patrón correcto partiendo de uno distorsionado. Sin embargo, está creciendo el uso de redes neuronales en distintos tipos de sistemas de control.

Desde el punto de vista de los casos de aplicación, la ventaja de las redes neuronales reside en el procesado paralelo, adaptativo y no lineal.

El dominio de aplicación de las redes neuronales también se lo puede clasificar de la siguiente forma: asociación y clasificación, regeneración de patrones, regresión y generalización, y optimización.

7.2.1 Asociación y clasificación

En esta aplicación, los patrones de entrada estáticos o señales temporales deben ser clasificados o reconocidos. Idealmente, un clasificador debería ser entrenado para que cuando se le presente una versión distorsionada ligeramente del patrón, pueda ser reconocida correctamente sin problemas. De la misma forma, la red debería presentar cierta inmunidad contra el ruido, esto es, debería ser capaz de recuperar una señal "limpia" de ambientes o canales ruidosos. Esto es fundamental en las aplicaciones holográficas, asociativas o regenerativas.

- Asociación: de especial interés son las dos clases de asociación: autoasociación y heteroasociación. Como ya se mencionó en el apartado 6.8, el problema de la autoasociación es recuperar un patrón enteramente, dada una información parcial del patrón deseado. La heteroasociación es recuperar un conjunto de patrones B, dado un patrón de ese conjunto. Los pesos en las redes asociativas son a menudo predeterminados basados en

la regla de Hebb. Normalmente, la autocorrelación del conjunto de patrones almacenado determina los pesos en las redes autoasociativas. Por otro lado, la correlación cruzada de muchas parejas de patrones se usa para determinar los pesos de la red de heteroasociación.

- Clasificación no Supervisada: para esta aplicación, los pesos sinápticos de la red son entrenados por la regla de aprendizaje no supervisado, esto es, la red adapta los pesos y verifica el resultado basándose únicamente en los patrones de entrada.

- Clasificación Supervisada: esta clasificación adopta algunas formas del criterio de interpolación o aproximación. En muchas aplicaciones de clasificación, por ejemplo, reconocimiento de voz, los datos de entrenamiento consisten de pares de patrones de entrada y salida. En este caso, es conveniente adoptar las redes Supervisadas, como las bien conocidas y estudiadas redes de retropropagación. Este tipo de redes son apropiadas para las aplicaciones que tienen una gran cantidad de clases con límites de separación complejos.

7.2.2 Regeneración de patrones

En muchos problemas de clasificación, una cuestión a solucionar es la recuperación de información, esto es, recuperar el patrón original dada solamente una información parcial. Hay dos clases de problemas: temporales y estáticos. El uso apropiado de la información contextual es la llave para tener éxito en el reconocimiento.

7.2.3 Regeneración y generalización

El objetivo de la generalización es dar una respuesta correcta a la salida para un estímulo de entrada que no ha sido entrenado con anterioridad. El sistema debe inducir la característica saliente del estímulo a la entrada y detectar la regularidad. Tal habilidad para el descubrimiento de esa regularidad es crítica en muchas aplicaciones. Esto hace que el sistema funcione eficazmente en todo el espacio, incluso cuando ha sido entrenado por un conjunto limitado de ejemplos.

7.2.4 Optimización

Las Redes Neuronales son herramientas interesantes para la optimización de aplicaciones, que normalmente implican la búsqueda del mínimo absoluto de una función de energía. Para algunas aplicaciones, la función de energía es fácilmente deducible; pero en otras, sin embargo, se obtiene de ciertos criterios de coste y limitaciones especiales.

7.3 Casos concretos de aplicación

A continuación se describe un caso concreto de aplicación de redes neuronales.

Planificación del staff de empleados.

Hoy más que nunca, las empresas están sujetas a la presión de los elevados costos. Esto puede verse en diferentes sectores corporativos, tales como la planificación del staff de empleados. Desde el punto de vista de las empresas, un empleado que falla al ejecutar la mayor parte de las tareas asignadas, evidencia una baja productividad. Por el otro lado, esta situación es frustrante para el empleado. Ambos efectos causan costos, los cuales podrían evitarse realizando antes una prueba de aptitud. Estos problemas no solamente son originados por los empleados nuevos, sino también por aquellos que son reubicados dentro de la misma empresa.

En este proyecto de investigación se examinó hasta donde la predicción de aptitudes puede llevarse a cabo por una red neuronal, cuya topología suministre una tarea satisfactoria y así lograr una predicción más exitosa.

Base de datos y codificación:

La base de datos inicial contenía información resultante de una investigación que realizaron por medio de un cuestionario. Las respuestas obtenidas a través del mismo las utilizaron para acumular información acerca de las cualidades específicas y habilidades técnicas de cada individuo del personal indagado. Para cada pregunta, les fue posible categorizar la respuesta en un intervalo que va de 1 a 5; constituyendo así la entrada que presentaron a la red neuronal. Al entrevistado, posteriormente, lo examinaron en el orden de obtener una cifra representativa de sus aptitudes. De esta manera el conjunto de datos de entrenamiento quedó formado de la siguiente forma:

- Respuesta obtenidas a través del cuestionario = datos de entrada.
- Cifra representativa de la aptitud de la persona encuestada = salida deseada.

El primer problema que se les presentó fue cómo codificar los datos obtenidos, decidiendo transformarlos dentro del intervalo [0.1, 1.0].

Cómo codificar la salida objetivo fue la próxima pregunta que consideraron. Normalmente la compañía sólo quiere conocer si una persona ejecutará bien o mal la tarea determinada, o si su desempeño será muy bueno, bueno, promedio, malo o muy malo. Consecuentemente, (*a*) asignaron la salida dada dentro de varias clases y (*b*) transformaron las cifras representativas dentro del intervalo [0, 1], utilizando en parte una función lineal.

Algoritmo de aprendizaje:

Ensayaron diferentes algoritmos de aprendizaje, de los cuales dos fueron escogidos como los más apropiados: Propagación Rápida (Quickpropagation) y Propagación Elástica (Resilient Propagation).

- Quickpropagation: es una modificación del algoritmo estándar de backpropagation. A diferencia de este, la adaptación de los pesos no es solamente influenciada por la sensibilidad actual, sino también por la inclusión del error previo calculado.
- Resilient Propagation: es otra modificación del algoritmo estándar de backpropagation. En oposición a este, la adaptación de los pesos es influenciada por el signo de la sensibilidad actual y antecesora, y no por su cantidad.

Topología de la red:

Evaluaron diferentes topologías de redes, las cuales no serán detalladas. La pregunta fue: (*a*) ¿cuántas capas ocultas son necesarias?, (*b*) ¿cuántas neuronas en cada una de ellas? La primera prueba que hicieron mostró que para este propósito la red debía contener 2 capas ocultas, con la primera formada por tantas neuronas como la capa de entrada y la segunda por un número menor que la primera (exactamente la mitad como mucho).

Resultados obtenidos a partir de los ensayos:

El primer resultado que consiguieron al intentar predecir la cifra representativa correcta fue relativamente mala. Asumieron que esto fue causado por el hecho de que el número de neuronas de entrada en proporción al número de ejemplos dados en el conjunto de datos de entrenamiento fue elevado. La pequeña base de datos, conforme con la gran capa de entrada, fue suficiente para realizar una tosca predicción, pero no para dar la correcta cifra representativa.

Lo mencionado en el párrafo anterior hizo que enfocaran toda la atención en reducir el número de neuronas de entradas en forma apropiada. También examinaron la red con la cual se logró el mejor resultado, en función de conseguir indicadores de las entradas que demostraran ser importantes y cuales no.

Entonces, reduciendo el número de neuronas de entrada y formando nuevas redes, consiguieron un resultado bastante bueno para la predicción de las clases y aún para la predicción de la cifra representativa correcta.

En otra serie de test, examinaron los resultados que podrían favorecer a un mejoramiento por agrupación de las neuronas de entrada para las preguntas interdependientes. Cada grupo, que representaba una habilidad especial, fue conectado exactamente a una neurona en la primera capa oculta. La razón para esto fue que haciendo ciertas conexiones se reduce beneficiosamente el espacio de búsqueda, si y solo si, las conexiones representan la estructura correcta, pero puede reducir el espacio de búsqueda inapropiadamente por prohibición de otras conexiones.

8. Herramientas de minería de datos

Existen infinidades de herramientas de Data Mining para el análisis de datos en gran dimensión, abarcando desde herramientas propietarias hasta las Open Source. A continuación mencionaremos algunas de estas herramientas mas utilizadas para el análisis de datos[24-26]:

8.1 Herramientas propietarias

8.1.1 IBM / SPSS Statistics

Herramienta de data mining que permite desarrollar modelos predictivos y desplegarlos para mejorar la toma de decisiones. Está diseñada teniendo en cuenta a los usuarios empresariales, de manera que no es preciso ser un experto en data mining.

Figura 14: Logotipo IBM SPSS

8.1.2 MicroStrategy Data Mining Services /Microstrategy:

Componente de la plataforma de BI de MicroStrategy que proporciona a los usuarios, modelos predictivos de data mining. Permite realizar tareas de data mining mediante el uso de métricas construidas con funciones predictivas o importadas de modelos de datos de herramientas de data mining de terceros.

Figura 15: Logotipo MicroStrategy

8.1.3 Microsoft SQL Server 2008 Datamining

Solución que ofrece un entorno integrado para crear modelos de minería de datos (Data Mining) y trabajar con ellos. La solución SQL Server Data Mining permite el acceso a la información necesaria para tomar decisiones inteligentes sobre problemas empresariales complejos. Data Mining es la tecnología de BI que ayuda a construir modelos analíticos complejos e integrar esos modelos con sus operaciones comerciales.

Figura 16: Logotipo SQL Server

8.1.4 SAS Enterprise Miner / SAS

Solución de minería de datos que proporciona gran cantidad de modelos y de alternativas. Permite determinar pautas y tendencias, explica resultados conocidos e identifica factores que permiten asegurar efectos deseados. Además, compara los resultados de las distintas técnicas de modelización, tanto en términos estadísticos como de negocio, dentro de un marco sencillo y fácil de interpretar.

Figura 17: Logotipo SAS

8.1.5 Oracle DataMining

Función de Oracle 11g Enterprise Edition que permite diseñar aplicaciones de BI que más tarde realizan funciones de “minería” en las bases de datos corporativas para descubrir nueva información e integrarla con las aplicaciones de negocio.

Figura 18: Logotipo Oracle

8.2 Herramientas Open Source

8.2.1 Orange

Es un componente de minería de datos y también es un software de aprendizaje de máquina, que permite una programación visual, rápida y versátil para un análisis exploratorio de datos, aunque es una herramienta poderosa, sigue siendo amigable e intuitiva. También permite pre procesamiento, filtros de información, modelación de datos, evaluación de modelos y técnicas de exploración. Además, cuenta con componentes de aprendizaje propio, complementos de informática y minería de texto.

Figura 19: Logotipo Orange

8.2.2 Rapid Miner (Anteriormente conocido como YALE)

Con un lenguaje de programación Java, RadipMiner permite realizar un análisis avanzado de los datos, a través de plantillas. Esta herramienta ofrece un servicio excelente, ocupando una de las primeras posiciones entre las mejores herramientas de data mining. Es un entorno para aprendizaje mecánico y experimentos de data mining, utilizado tanto en investigación como en tareas de día a día por diferentes empresas. Produce sus resultados en archivos XML y cuentan con la interface gráfica del mismo programa. Provee más de 500 operadores para los principales procesos de aprendizaje en máquina y al tiempo

combina esquemas y atributos de evaluación. Además, dispone de la funcionalidad de pre-procesamiento y visualización de datos, análisis predictivo y modelos estadísticos, así como evaluación y despliegue de la información.

Figura 20: Logotipo RapidMiner

8.2.3 Weka

(WEKA: Waikato Environment for Knowledge Analysis) Maneja lenguaje Java, es una de las herramientas para aplicación de tareas de data mining más reconocidas. Es una de las herramientas para aplicación de tareas de data mining más reconocidas, que permite proceso previo, clustering o generación de grupos de datos, clasificación, regresiones, visualización y selección de propiedades. Una de las propiedades más interesantes de este software, es su facilidad para añadir extensiones, modificar métodos, entre otros.

Características de Weka:

- Está disponible libremente bajo la licencia pública general GNU.
- Es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
- Es fácil de usar por un principiante gracias a su interfaz gráfica de usuario.
- Contiene una extensa colección de técnicas para pre procesamiento de datos y modelado.

Figura 21: Logotipo Weka

8.2.4 jHepWork

Diseñado para científicos, ingenieros y estudiantes, es una herramienta gratuita y de uso libre, que permite el análisis de datos mediante la creación de un entorno comprensible, amigable y adaptable a programas comerciales. Contiene librerías científicas en Java para funciones matemáticas, y algoritmos de minería de datos. Esta herramienta es un poco más avanzada y se requiere más alto conocimiento, el lenguaje usado es Jython, aunque también funciona a la perfección en Java.

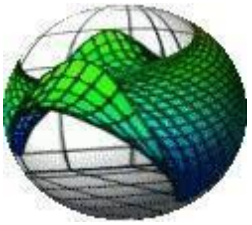

Figura 22: Logotipo jHepWork

8.2.5 KNIME

Konstanz Information Miner, es un software de integración de datos amigable, intuitivo y fácil de usar, que permite el procesamiento, análisis y exploración de datos, desde su plataforma. Le permitirá crear visualmente flujos de datos, ejecutar análisis selectivamente, estudiar los resultados, modelar y generar vistas interactivas, para facilitar la toma de decisiones a nivel gerencial. KNIME está escrito en Java y está basado en Eclipse, en el site de Kmine encontrará extensiones o plugins adicionales que proporcionando así una funcionalidad adicional que permitirá a los usuarios añadir módulos de texto, imagen, procesamiento de series de tiempo y la integración de varios otros proyectos de código abierto, tales como el lenguaje de programación de R, WEKA, el Kit de desarrollo de la Química, y LIBSVM.

Figura 23: Logotipo KNIME

8.2.6 R - DM

Es un lenguaje de Script para manipulación de datos análisis, estadísticos y visualización con base en el respetado lenguaje C. Es comparable a menudo en poder de productos comerciales. Disponible en Windows Mac y Linux. Facilidad de uso y extensibilidad ha elevado la popularidad de R sustancialmente en los últimos años. Además de lo mencionado proporciona técnicas estadísticas y gráficos, incluyendo lineal y no lineal de modelado, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupación, y otros.

Figura 24: Logotipo R

8.3 Herramientas más utilizadas en los últimos años[27-32]

Rexer Analytics viene llevando a cabo desde el 2007 y de manera regular encuestas sobre la elección del software que utilizan los profesionales del data mining y el analytics en general, y los resultados de la Encuesta 2013 de Rexer Analytics a los data miners, fueron presentados en la conferencia Predictive Analytics World realizada en Boston en setiembre del 2013. A continuación mostramos lo más destacado de las respuestas hechas por 1039 participantes no vendedores. (Los resultados completos pueden ser obtenidos solicitándolos a Rexer Analytics).

R es la herramienta más popular para data mining, la cual es utilizada al menos de forma ocasional por el 70% de los encuestados. Esta popularidad se mantiene en todos los grupos de encuestados: En el 2013 R fue la herramienta más utilizada entre los profesionales de data mining corporativos (70%), los consultores de data mining (73%), los académicos de data mining (75%) y los data miners que pertenecen a entidades del gobierno y a las organizaciones no gubernamentales (67%). Y mientras que los data miners reportan en promedio el uso de 5 herramientas de software para sus tareas de explotación de datos, R es considerada como la herramienta principal en la encuesta, con un 24%.

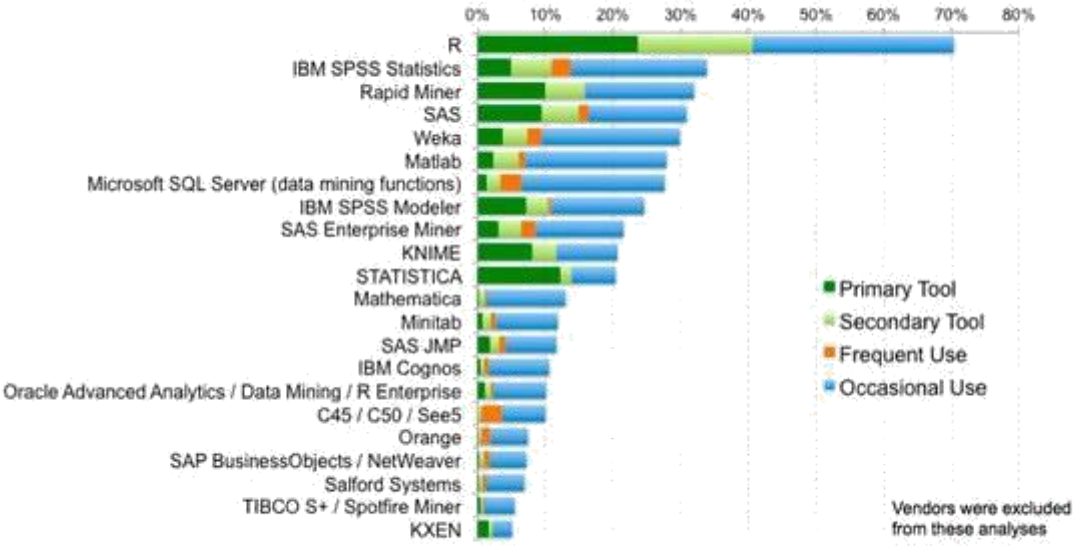

Figura 25: Encuesta 2013, Rexer Analytics

La popularidad de R en el 2013 se disparó, y su predominio aumentó en todas las encuestas que Rexer ha realizado desde el 2007, tanto en su uso general como su elección como herramienta principal.

Uso de R por profesionales de data mining a través de los años

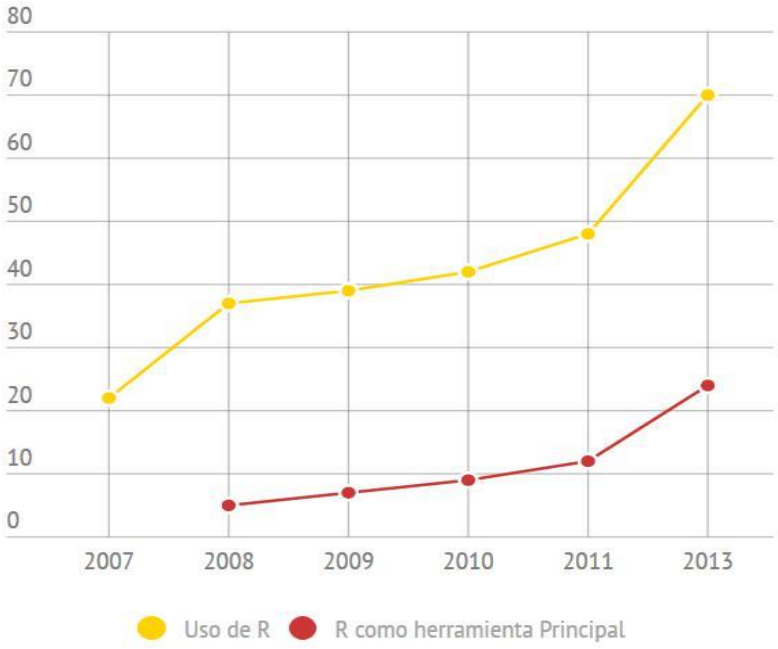

Figura 26: Uso de R por profesionales de DM a través de los años, Rexer Analytics

Según la encuesta realizada por Rexer Analytics, el 70% de los profesionales de data mining usan R para sus tareas de explotación de datos, mientras que el 24% de ellos respondieron que R es la herramienta principal que utilizan.

La mayoría de los usuarios se sienten satisfechos con R, con más del 85% de usuarios que contestaron sentirse satisfechos y muy satisfechos con esta herramienta. La plataforma de software con la mayor cantidad de usuarios insatisfechos son SAS y SAS Enterprise Miner.

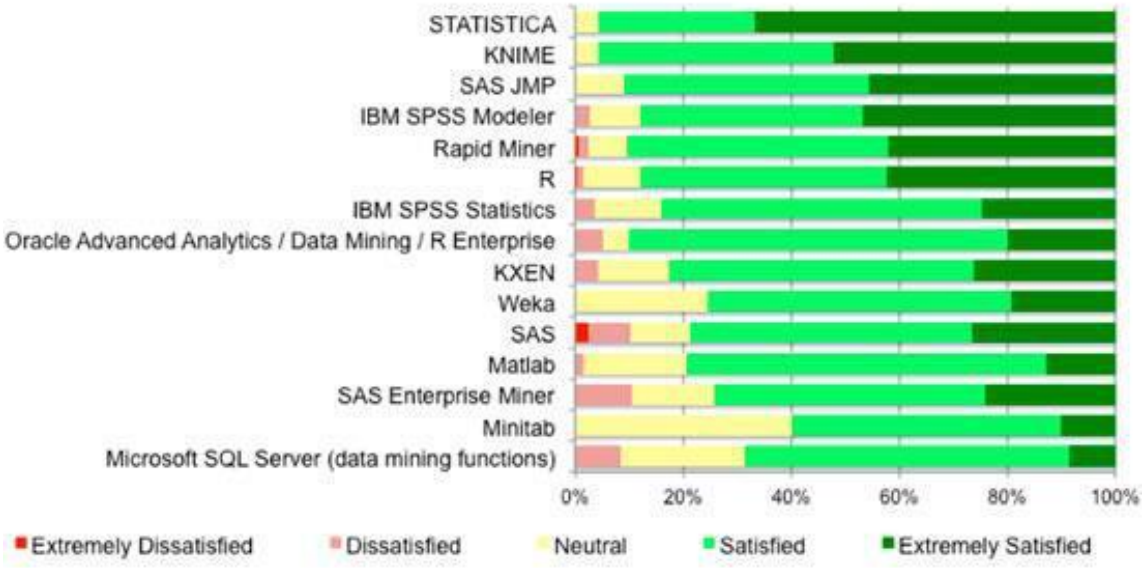

Figura 27: Resultado encuesta uso R, Rexer Analytics

La encuesta anual de KDNuggets de los lenguajes top para analytics, data mining y la ciencia de los datos, lanzada en Setiembre del 2013, nombra a R como el software más popular por tercer año consecutivo.

Ya en el 2014, KDNuggets realizó una encuesta a sus lectores preguntando lo siguiente: “¿Qué lenguajes de programación estadísticas se utiliza para un trabajo de minería de datos para el análisis de datos en el año 2014?”. Una vez más, R era la respuesta N° 1. (R era también la respuesta N° 1 en las encuestas similares en 2013, 2012 y 2011.) Los 5 mejores selecciones de los 719 encuestados eran:

1. R (352 encuestados)
2. SAS (262)
3. Python (252)
4. SQL (220)
5. Java (89)

Los encuestados fueron capaces de seleccionar varios lenguajes, por lo que los totales suman más que el número de los encuestados. De hecho, el análisis de software de datos utilizados juntos es muy interesante. En cuanto a la parte superior, hay un poco de coincidencia entre los usuarios de Python y R, pero poco cruce entre los usuarios de SAS.

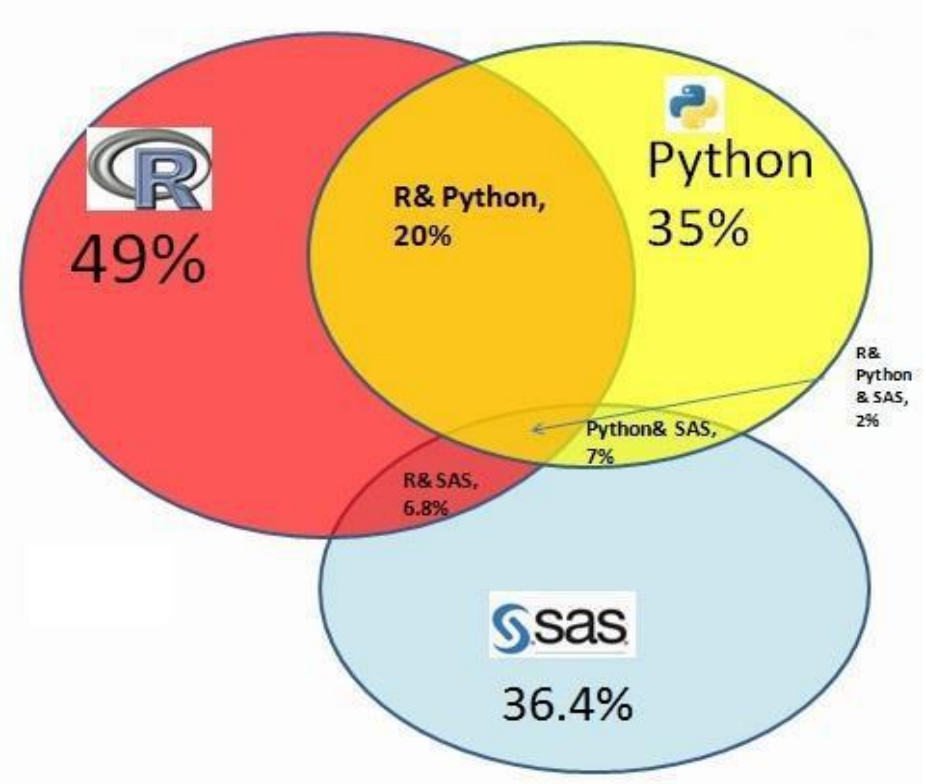

Figura 28: Lenguajes más usados para el análisis DM 2014, KDNuggets

R también ha altamente clasificado en una serie de otros sondeos y encuestas, las cuales se pueden ver en nuestra etiqueta de popularidad R. Los resultados completos y un análisis más detallado de esta encuesta KDNuggets.

La 16.ª encuesta anual de software KDnuggets continuó consiguiendo gran atención por parte de la analítica y la comunidad de minería de datos y proveedores, que atrae a cerca de 3.000 votantes, que eligieron a partir de un número récord de 93 herramientas diferentes.

R es la herramienta global más popular entre los mineros de datos, aunque el uso de Python está creciendo rápidamente. RapidMiner sigue siendo la suite más popular para la minería de datos ciencia / datos. También observamos un gran aumento en el uso de herramientas Hadoop / Big Data (29%, frente al 17% en 2014), y el crecimiento más rápido es por Spark cuya cuota de uso creció más de 3 veces. Otras herramientas con fuerte crecimiento incluyen H2O (0xdata), Actian, MLlib, y Alteryx.

Aquí los resultados y el análisis de la encuesta de las 10 mejores herramientas por participación en el uso:

R conduce RapidMiner, Python se pone al día, las herramientas de grandes volúmenes de datos crecen, se inflama Spark

La participación por regiones fueron: Estados Unidos / Canadá (41,5%), Europa (38,4%), Asia (8,2%), América Latina (6,3%), Australia / Nueva Zelanda (3,1%), África / Medio Oriente (2,5%).

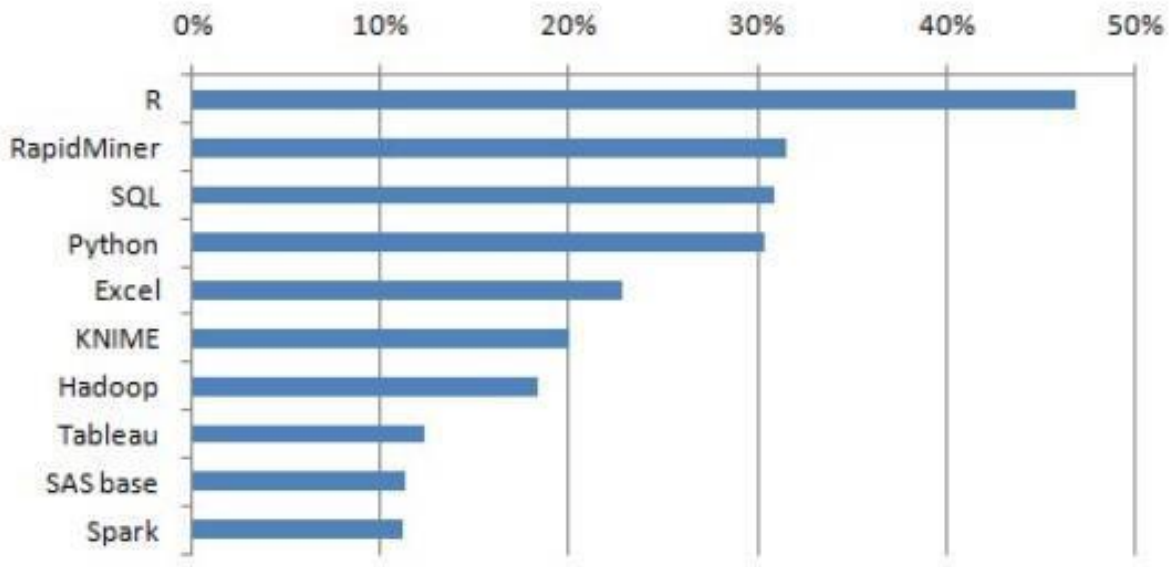

Figura 29: Top 10 analytics data mining 2015, KDNuggets

Las 10 mejores herramientas por parte de los usuarios eran:

1. R, 46.9% (38.5% in 2014)
2. RapidMiner, 31.5% (44.2% in 2014)
3. SQL, 30.9% (25.3% in 2014)
4. Python, 30.3% (19.5% in 2014)
5. Excel, 22.9% (25.8% in 2014)
6. KNIME, 20.0% (15.0% in 2014)
7. Hadoop, 18.4% (12.7% in 2014)
8. Tableau, 12.4% (9.1% in 2014)
9. SAS, 11.3 (10.9% in 2014)
10. Spark, 11.3% (2.6% in 2014)

En comparación a 2014 Analytics/Encuesta de software de data mining, Tableau y Spark recién llegaron al top 10, desplazando Weka y Microsoft SQL Server. La distinción entre el software comercial y libre es cada vez más difícil de realizar, con muchas herramientas que tiene tanto una versión gratuita / comunidad y la versión comercial / empresarial. Hemos clasificado cada herramienta de acuerdo con el tipo primario de la última versión, así que pusimos RapidMiner en la categoría comercial y KNIME en la categoría de software libre.

En el 2015, el 91% de los votantes utiliza software comercial y el 73% utiliza el software libre. Aproximadamente el 27% utiliza sólo el software comercial, y sólo el 9% utiliza el software libre. Por primera vez, una mayoría del 64% utiliza el software libre y comercial, por encima del 49% en 2014 como se muestra en la figura 29.

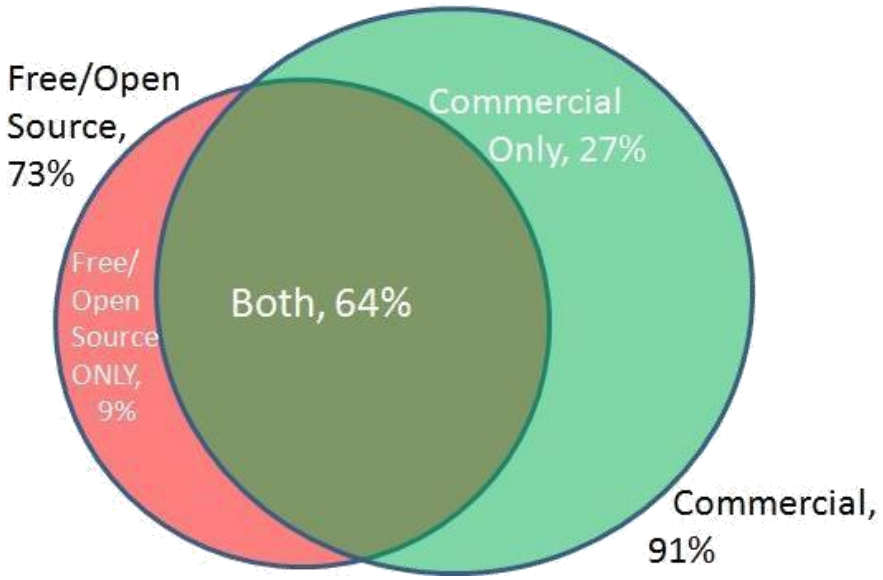

Figura 30: Analytics data mining 2015, KDNuggets

9. Mitos y errores de la minería de datos[33]

9.1 Mitos de la minería de datos

Mito N° 1: La minería de datos se realizó en el laboratorio, por un experto en tecnología

La minería de datos utiliza la tecnología avanzada, y su trabajo, en particular los de las técnicas de modelado, es improbable que sean comprendidos por la más amplia comunidad de IT. ¿Significa esto que la minería de datos debería tener lugar en el laboratorio y se realizará sólo por aquellos que entienden todos los matices de la tecnología que está en juego?

Es verdad, porque la minería de datos es un proceso de negocio en el que el conocimiento de negocio es de vital importancia: el valor de la minería de datos se realiza sólo cuando los resultados se ponen en uso en operaciones comerciales.

Cuando se realiza sin conocimiento del negocio, la minería de datos puede producir resultados sin sentido, lo que es fundamental que la minería de datos se realizará por alguien con un amplio conocimiento del problema de negocio. Muy rara vez es esta la misma persona que tiene un amplio conocimiento de la tecnología de minería de datos. Es la responsabilidad de los proveedores de

instrumentos de minería de datos para asegurar que los instrumentos son accesibles a los usuarios de negocios.

De igual importancia es la necesidad de desplegar los resultados en el negocio, para darles un uso. Los mineros de datos deben planificar en el inicio de un proyecto de cómo sus resultados se ajustan a los procesos de negocio operativos. Las organizaciones deben adquirir la infraestructura que les permite desplegar los resultados de la minería de datos de manera eficiente a través de la organización, y los proveedores de instrumentos deben garantizar que sus instrumentos se adaptan fácilmente a esta infraestructura.

Mito N° 2: La minería de datos tiene que ver con algoritmos

Una persona de negocios que asiste a una típica conferencia de minería de datos o lee sus actuaciones podría constituir la impresión de que la minería de datos tiene que ver con los algoritmos de análisis de datos avanzados. Esta concepción errónea puede resumirse como tal: “Todo lo que necesitas para la minería de datos son buenos algoritmos. Cuanto mejor sean tus algoritmos, mejor será tu minería de datos. Mejorar la eficacia de la extracción de datos para mejorar nuestro conocimiento de los algoritmos”.

Tener este punto de vista es no entender el proceso de minería de datos. La minería de datos es un proceso compuesto de muchos elementos, como la formulación de los objetivos de negocio, la cartografía de los objetivos empresariales con los objetivos de la minería de datos, la adquisición, comprensión, y pre-tratamiento de los datos, evaluación y presentación de los resultados de análisis, y la implementación de estos resultados para lograr beneficios de los negocios.

Esto no es minimizar la importancia de los algoritmos de la minería de datos nuevos o mejorados.

El problema se produce cuando los datos de los mineros se centran demasiado en los algoritmos y pasan por alto el otro 90-95 por ciento del proceso de minería de datos.

Las consecuencias de este error puede ser desastroso para un proyecto de minería de datos, posiblemente resultando en una incapacidad para producir resultados útiles. Mineros de datos experimentados reconocen la necesidad de una visión más amplia del proceso de minería de datos.

Mito N° 3: La minería de datos es todo acerca de la exactitud de predicción

Si bien la minería de datos no es *todo* acerca de los algoritmos de análisis de datos, hay una parte de la minería de datos que se *trata* de algoritmos. Esto plantea la pregunta, "¿Cómo se puede juzgar la calidad de un algoritmo?"

Se podría pensar que el criterio principal será la exactitud predictiva de los modelos que genera. Esta opinión, sin embargo, tergiversa el papel de los algoritmos en el proceso de minería de datos.

Es cierto que un modelo de predicción debe tener algún grado de exactitud, porque esta demuestra que realmente ha descubierto patrones en los datos. Sin embargo, la utilidad de un algoritmo o el modelo también está determinada por una serie de otras propiedades, una de las cuales es la comprensión del modelo resultante exige un conocimiento profundo técnico o es algo que pueda ser entendido por un analista de negocio típico.

Los mineros de datos que creen que la precisión de las predicciones es el principal criterio de evaluación del algoritmo pueden utilizar algoritmos que sólo pueden ser utilizados por expertos en tecnología. Estos algoritmos jugarán sólo el papel más limitado, ya que la minería de datos es un proceso que es conducido por la experiencia empresarial; sino que descansa en la aportación y participación de los profesionales de la empresa no-técnicos con el fin de tener éxito.

Mito N° 4: La minería de datos requiere de un almacén de datos

Los empresarios suelen pensar que un almacén de datos es un requisito previo para la minería de datos. Es una idea errónea sutil acerca de la relación entre las dos tecnologías.

Es cierto que la minería de datos puede beneficiarse de un almacén de datos que está bien organizado, relativamente limpia, y de fácil acceso. Esto es especialmente cierto si el depósito ha sido construido con la minería de datos específicamente en mente y con el conocimiento de los requisitos del proyecto de minería de datos. Si esto no ha sido el caso, sin embargo, los datos almacenados pueden ser menos útiles para la extracción de datos de la fuente o los datos operativos. En el peor de los casos, los datos almacenados pueden ser completamente inútiles (por ejemplo, si sólo un resumen de los datos se almacenó).

Una representación más precisa de la relación entre los dos sería que la minería de datos aproveche de un almacén de datos bien diseñado; y que la construcción de dicho depósito aproveche siempre del primero haciendo algunas exploraciones de minería de datos.

Mito N° 5: La minería de datos es todo a propósito de las cantidades enormes de datos

Las explicaciones al principio de la minería de datos a menudo se inició con declaraciones como, “Ahora recogemos más datos que nunca, pero ¿cómo vamos a beneficiarse de estos grandes almacenes de datos?” Centrarse en el tamaño de los almacenes de datos ha proporcionado una introducción práctica al tema de la minería de datos, pero sutilmente ha tergiversado su naturaleza.

Si bien hay muchos grandes conjuntos de datos que las organizaciones pueden beneficiarse de la minería, sería un error creer que estos deben ser el único foco de la minería de datos. Muchos proyectos de minería de datos útiles se realizan en las pequeñas o medianas series de datos algunas, por ejemplo, contienen sólo unos pocos cientos o miles de registros.

Suscribiendo a la creencia errónea de que la minería de datos sólo es apropiado para grandes almacenes de datos daría lugar a las organizaciones elegir los instrumentos que sacrifican la convivencia para la evolutividad cuando, de hecho,

ambos atributos son esenciales. Para citar a un cliente de uno de los principales instrumentos de minería de datos: “Otros instrumentos de minería de datos optimiza el tiempo de la máquina, pero esto instrumento optimiza mi tiempo” ya sea que los conjuntos de datos son grandes o pequeños, las organizaciones

9.2 Errores de la minería de datos

Error N° 1: Sepultado bajo montañas de datos

La minería de datos debe ser un proceso interactivo e iterativo, en el que el analista aplica el conocimiento del negocio importante y está “comprometido” con los datos y el problema de negocio. Sin embargo, aquellos que mantienen con el mito N° 5 (que la minería de datos es de grandes cantidades de datos) suelen suponer que este proceso debe ser aplicado a todos los datos disponibles.

Esto puede llevar a los intentos de minar volúmenes de datos para que el hardware y software disponibles no pueden proporcionar una respuesta interactiva disponible. En estas situaciones, el proceso de minería de datos se hace lento, y cuando una pregunta se contesta, el analista no puede recordar por qué se preguntó.

La manera de evitar este escollo es emplear alguna forma de muestreo. Por ejemplo, si tenemos un millón de clientes y una deserción (o “churn”) del 20 por ciento anual, no tenemos que parcelar nuestros gráficos o construir nuestros modelos con el pleno millones de ejemplos, o incluso 500,000.

Considere las siguientes preguntas y respuestas:

P: ¿Cuántos perfiles de “churn” que esperamos encontrar?
R: Tal vez diez
P: ¿Cuántos ejemplos de cada perfil que necesitamos para construir un modelo exacto?
R: Tal vez un millón

Por lo tanto, una muestra de diez o veinte mil “churners” y un número equivalente de no-“churners” es probable que sea suficiente para este análisis.

Tenga en cuenta que esto no significa que los mineros de datos no se encontraran con la necesidad de construir modelos de millones de ejemplos, sólo que no tienen que asumir que deben hacerlo, sólo porque los datos están disponibles.

Error N° 2: La Misteriosa desaparición de terabyte

Este es un fenómeno común, pero no siempre es una trampa. Se refiere al hecho de que, para un determinado problema de minería de datos, la cantidad de datos disponibles y pertinentes puede ser mucho menos de lo inicialmente supuesto.
Considere el siguiente escenario: Usted es un consultor de minería de datos, y su cliente es un gran banco, que desea minar los datos de sus clientes para determinar el riesgo de crédito. El banco tiene terabytes de datos sobre sus clientes y le preocupa que los recursos informáticos disponibles pueden ser inadecuadas para minar este volumen de datos.

Aquí es cómo la situación podría desarrollarse. Los diferentes tipos de crédito (préstamos personales, préstamos comerciales, los descubiertos) presentan diferentes patrones de riesgo de crédito, de modo que cada proyecto de minería de datos se concentrará en un solo tipo de prestatario. Expertos en el dominio del banco juzgan una serie de factores a ser relevante, y el banco, planificando el futuro, comenzó a recoger datos sobre estos factores hace unos 18 meses.

Desde entonces, casi un millón de casos de deudas incobrables del tipo de referencia se han producido. Así, los datos de referencia consiste en menos de un millón de casos de malas deudas y una muestra de un suministro abundante de casos de buenas deudas-digamos 3,000 registros en total. Esto es probablemente lo suficiente, pero sólo lo suficiente, para construir un modelo de predicción precisa. De alguna manera, la necesidad de minar los terabytes de los datos ha “misteriosamente” desaparecido, no del todo teniendo el minero de datos con él (esta vez).

Error N° 3: Minería de datos desorganizada

La minería de datos pueden a veces, a pesar de la mejor de las intenciones, tener lugar de manera ad hoc, sin objetivos claros y sin idea de cómo los resultados serán utilizados. Esto conduce a la pérdida de tiempo y los resultades inutilizables.

Para producir resultados útiles, es fundamental contar con objetivos de negocio claramente definidos, los objetivos de la minería de datos y los planes de despliegue, todos formulados en el principio del proyecto. Una manera sencilla de asegurar esto es utilizar un proceso estándar, como el Cross-Industry Standard Process para la Minería de Datos (CRISP-DM). Este proceso garantiza la correcta preparación para la minería de datos y proporciona un lenguaje común para la comunicación de los métodos y resultados. Los instrumentos de minería de datos deben apoyar los modelos de procesos estándar.

Error N° 4: Conocimiento del negocio insuficiente

He mencionado el papel crucial que desempeña el conocimiento del negocio en la minería de datos. Sin él, las organizaciones no pueden lograr resultados útiles ni guiar el proceso de minería de datos hacía ellos.

A veces se supone que el usuario final puede razonablemente decir a los mineros de datos: “Estos son los datos, por favor vaya, haga su minería de datos, y regresa con la respuesta”. Si esto llegara a suceder, el proyecto, en el mejor, toma muchas iteraciones largas y costosas para producir resultados útiles. En el peor, los resultados serían un desorden, y el proyecto fracasaría.

Esta trampa sólo puede ser evitada mediante la participación, en cada etapa del proceso de minería de datos, tanto con el usuario final y alguien con un conocimiento detallado del negocio. Idealmente, el minero de datos o un consultor de minería de datos tendría el conocimiento del negocio. Careciendo de ello, el minero de datos, literalmente, debe sentarse al lado de alguien con el conocimiento del negocio requerido que entiende la cuestión en examen. Para que este trabajo sea eficaz, un medio ambiente de minería de datos altamente interactivo con buen tiempo de respuesta es necesario.

Error N° 5: El conocimiento de datos insuficiente

Para llevar a cabo la extracción de datos, debemos ser capaces de responder a las preguntas como “¿Qué significan los códigos en este campo?” y “¿Puede haber más de un registro por cliente en esta tabla?”. En algunos casos, esta información es sorprendentemente difícil de encontrar. Puede ser que el experto de datos ha dejado la organización o trasladado a otro departamento o, en el caso de los sistemas de legado, no puede ser ningún experto de datos del todo.

Este problema se agrava cuando se subcontrata la base de datos o la gestión de almacenes de datos: el proveedor externo es aún menos motivado que la organización de usuarios de mantener esta información “sólo en caso de que pudieran ser necesarios en el futuro”.

No hay ninguna resolución simple a este problema. Los departamentos de IT deben ser conscientes de la necesidad de mantener la información sobre las bases de datos de su organización. Además, cuando se propone un proyecto de minería de datos, los mineros de datos deben considerar la cantidad disponible de conocimiento de datos y evaluar los riesgos causados por su ausencia o escasez.

Error N° 6: Suposiciones erróneas, cortesía de los expertos

Empresas y expertos de datos son recursos cruciales, pero esto no significa que el minero de datos debe aceptar incondicionalmente todas las declaraciones que hacen. El minero de datos debe tratar de confirmar la validez de las declaraciones de los expertos.

Los ejemplos típicos de las declaraciones erróneas o engañosas pueden ser:

- Ningún cliente puede tener cuentas en ambos tipos.
- Ningún caso incluirá más de un evento de este tipo.
- Sólo los códigos siguientes estarán presentes en este campo.

Los mineros de datos deben verificar estas declaraciones mediante el examen de los datos. Esto es particularmente importante porque el tratamiento de los datos dependerá de su exactitud. Idealmente, los errores en las suposiciones acerca de los datos pueden ser detectados antes de dar lugar a errores en el tratamiento de los datos.

10. Aplicaciones y análisis de casos

10.1 Empresariales

Detección de fraudes en las tarjetas de crédito

Problemática: Hubo un tiempo que las instituciones financieras a escala mundial perdieron del orden de 2.000 millones de dólares en fraudes cometidos con tarjetas de crédito. El *Falcon Fraud Manager* es un sistema inteligente que examina transacciones, propietarios de tarjetas y datos financieros para intentar detectar y paliar el número de fraudes. En un principio estaba pensado, en instituciones financieras de Norteamérica, para detectar fraudes en tarjetas de crédito. Sin embargo, actualmente se le han incorporado funcionalidades de análisis en las tarjetas comerciales, de combustibles y de débito. El sistema Falcon ha permitido ahorrar más de seiscientos millones de dólares al año y proteger aproximadamente más de cuatrocientos cincuenta millones de pagos con tarjeta en todo el mundo, aproximadamente el 65% de todas las transacciones con tarjeta de crédito.

Resultado: La solución de Falcon usa una sofisticada combinación de modelos de redes neuronales para analizar el pago mediante tarjeta y detectar los más remotos casos de fraude. Lleva siendo usado durante más de 15 años y monitoriza alrededor de 450 millones de cuentas distribuidas en los 6 continentes.

Migración de clientes entre distintas compañías

Problemática: La migración de clientes de una operadora de comunicaciones móvil a otra. Este estudio fue desarrollado en una operadora española que básicamente situó sus objetivos en dos puntos:

- El análisis del perfil de los clientes que se dan de baja
- La predicción del comportamiento de sus nuevos clientes

Se analizaron las diferencias históricas entre clientes que habían abandonado la operadora (12,6%) y de los clientes que continuaban con su servicio (87,4%). También se analizaron las variables personales de cada cliente (estado civil, edad, sexo, nacionalidad, etc.). De igual forma se estudiaron para cada cliente la morosidad, la frecuencia y el horario de uso del servicio, los descuentos y el porcentaje de llamadas locales, interprovinciales, internacionales y gratuitas.

Al contrario de lo que se podría pensar, los clientes que abandonaban la operadora generaban ganancias para la empresa; sin embargo, una de las conclusiones más importantes radicaba en el hecho de que los clientes que se daban de baja recibían pocas promociones y registraban un mayor número de incidencias respecto a la media.

Resultado: Como resultado de este estudio de minería de datos se recomendó a la operadora hacer un estudio sobre sus ofertas y analizar profundamente las incidencias recibidas por esos clientes. Al descubrir el perfil que presentaban, la operadora tuvo que diseñar un trato más personalizado para sus clientes actuales con esas características. Para poder predecir el comportamiento de sus nuevos clientes se diseñó un sistema de predicción basado en la cantidad de datos que se podía obtener de los nuevos clientes comparados con el comportamiento de clientes anteriores.

Supermercados Wal-mart

Problemática: Hace algunos años uno de estos supermercados se hizo la pregunta sobre qué productos se vendían con mayor frecuencia en compañía de los pañales. Pues bien, “minearon” la base de datos y encontraron que en asociación con los pañales se vendían muy frecuentemente las cervezas. Además, se dieron cuenta que ambos productos se vendían principalmente los viernes en la tarde y eran comprados por hombres con edades entre los 25 y 35 años de edad.

Después de cierto tiempo descubrieron la razón de este hallazgo. El caso es que los paquetes de pañales son voluminosos, y las esposas, que en muchos casos hacen la compra de la casa, dejan los pañales para que el esposo los compre. El esposo y padre, compraba los pañales especialmente los viernes, en compañía de las cervezas para el fin de semana.

Resultados: Como consecuencia de esto el supermercado puso la cerveza al lado de los pañales. El resultado fue que los padres que normalmente llegaban a comprar los pañales y la cerveza, compraron más cervezas, y los que antes no compraban cerveza, empezaron a comprarla por la proximidad de ésta con los pañales. Finalmente las ventas de cerveza se dispararon.

10.2 Universidad

Conociendo si los recién titulados de una universidad llevan a cabo actividades profesionales relacionadas con sus estudios.

Problemática: Se realizó un estudio sobre los recién titulados de la carrera de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Chihuahua II en Mexico. Se quería observar si los recién titulados se insertaban en actividades profesionales relacionadas con sus estudios y, en caso negativo, se buscaba saber el perfil que caracterizó a los ex-alumnos durante su estancia en la universidad. Se deseaba concluir si con los planes de estudio de la universidad y el rendimiento del alumno se hacía una buena inserción laboral o si existían otras variables que participaban en el proceso. Dentro de la información considerada estaba el sexo la edad, la escuela de procedencia, el desempeño académico, la zona económica donde tenía su vivienda y la actividad profesional, entre otras variables. Mediante la aplicación de conjuntos aproximados se descubrió que existían cuatro variables que determinaban la adecuada inserción laboral, que son citadas de acuerdo con su importancia:

1. Zona económica donde habitaba el estudiante
2. Colegio de donde provenía
3. Nota al ingresar
4. Promedio final al salir de la carrera

Resultados: A partir de estos resultados, la universidad obtuvo que las tres características más importantes no tenían relación con la universidad, y si de la economía de la zona donde provenía el estudiante. Por lo que podía plantearse nuevas soluciones de tipo socioeconómico, como becas en empresas u otras.

10.3 Investigación espacial

Proyecto SKYCAT

Problemática: Durante seis años, el Second Palomar Observatory Sky Survey (POSS-II) coleccionó tres terabytes de imágenes que contenían aproximadamente dos millones de objetos en el cielo. Tres mil fotografías fueron digitalizadas a una resolución de 16 bits por píxel con 23040 x 23040 píxeles por imagen. El objetivo era formar un catálogo de todos esos objetos. El sistema Sky

Image Cataloguing and Analysis Tool (SKYCAT) se basa en técnicas de agrupación (clustering) y árboles de decisión para poder clasificar los objetos en estrellas, planetas, sistemas, galaxias, etc. con una alta confiabilidad.

Resultados: Los resultados han ayudado a los astrónomos a descubrir dieciséis nuevos quásares (señales radiales lejanas) con corrimiento hacia el rojo que los incluye entre los objetos más lejanos del universo y, por consiguiente, más antiguos. Los quásares son fuentes de Rayos X, radiación ultravioleta, luz visible y también infrarroja; en otras palabras, la emisión de radiación de los quásares resulta intensa en todo el espectro electromagnético. Estos quásares son difíciles de encontrar y permiten saber más acerca de los orígenes del universo.

10.4 Deporte

A.C Milán

Problemática: El AC de Milán utiliza un sistema inteligente para prevenir lesiones. El club posee redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta. Esto ayuda a seleccionar el fichaje de un posible jugador o a alertar al médico del equipo de una posible lesión. El sistema, creado por *Computer Associates International*, es alimentado por datos de cada jugador, relacionados con su rendimiento, alimentación y respuesta a estímulos externos, que se obtienen y analizan cada quince días.

El jugador lleva a cabo determinadas actividades que son monitorizadas por veinticuatro sensores conectados al cuerpo y que transmiten señales de radio que posteriormente son almacenadas en una base de datos.

Actualmente el sistema dispone de 5000 casos registrados que permiten predecir alguna posible lesión. Con ello, el club intenta ahorrar dinero evitando comprar jugadores que presenten una alta probabilidad de lesión, lo que haría incluso renegociar su contrato. Por otra parte, el sistema pretende encontrar las diferencias entre las lesiones de atletas de ambos sexos, así como saber si una determinada lesión se relaciona con el estilo de juego de un país concreto donde se practica el fútbol.

NBA: Knicks de New York y Patrick Ewing

Problemática: Los equipos de la NBA también utilizan aplicaciones inteligentes para apoyar a su cuerpo de entrenadores. El *Advanced Scout* es un software que emplea técnicas de Data Mining y que han desarrollado investigadores de IBM para detectar patrones estadísticos y eventos extraños. Tiene una interfaz gráfica muy amigable orientada a un objetivo muy específico: analizar el juego de los equipos de la National Basketball Association (NBA).El software utiliza todos los registros guardados de cada evento en cada juego: pases, encestes, rebotes y doble marcaje (double team) a un jugador por el equipo contrario, entre otros. El objetivo es ayudar a los entrenadores a aislar eventos que no detectan cuando observan el juego en vivo o en película. Un resultado interesante fue uno hasta entonces no observado por los entrenadores de los Knicks de Nueva York. El doble marcaje a un jugador puede generalmente dar la oportunidad a otro jugador de encestar más fácilmente. Sin embargo, cuando los Bulls de Chicago jugaban contra los Knicks, se encontró que el porcentaje de encestes después de que al centro de los Knicks, Patrick Ewing, le hicieran doble marcaje era extremadamente bajo, indicando que los Knicks no reaccionaban correctamente a los dobles marcajes. Para saber el porqué, el cuerpo de entrenadores estudió cuidadosamente todas las películas de juegos contra Chicago. Observaron que los jugadores de Chicago rompían su doble marcaje muy rápido de tal forma que podían tapar al encestador libre de los Knicks antes de prepararse para efectuar su tiro. Con este conocimiento, los entrenadores crearon estrategias alternativas para tratar con el doble marcaje.

Resultados: La temporada pasada, IBM ofreció el Advanced Scout a la NBA, que se convirtió así en un patrocinador corporativo. La NBA dio a sus veintinueve equipos la oportunidad de aplicarlo. Dieciocho equipos lo están haciendo hasta el momento obteniendo descubrimientos interesantes.

10.5 Textos: Text Mining

Estudios recientes indican que la mayor parte de la toda la información de una compañía está almacenada en forma de documentos. Sin duda, este campo de estudio es muy complejo y de dimensiones enormes, por lo que técnicas como pueden ser la *categorización de texto*, el *procesamiento de lenguaje natural*, la extracción y recuperación de la información o el aprendizaje automático, entre otras, apoyan al text mining (minería de texto). En ocasiones se confunde el text mining con la recuperación de la información (Information Retrieval o IR). Ésta última consiste en la recuperación automática de documentos relevantes mediante indexaciones de textos, clasificación, categorización, etc.

Generalmente se utilizan palabras clave para encontrar una página relevante. En cambio, el text mining se refiere a examinar una colección de documentos y descubrir información no contenida en ningún documento individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo.

Medicina

Una aplicación muy popular del text mining es relatada en Hearst (1999). Don Swanson intenta extraer información derivada de colecciones de texto. Teniendo en cuenta que los expertos sólo pueden leer una pequeña parte de todo lo que se publica en su campo, y por lo general tampoco pueden tener en cuenta los Nuevos desarrollos que se suceden en otros campos relacionados, y teniendo en cuenta que la cantidad de nuevos documentos que se publican es cada vez mayor, la aplicación de la minería de datos en colecciones de texto va resultando más importante. Así, Swanson ha demostrado cómo cadenas de implicaciones causales dentro de la literatura médica pueden conducir a hipótesis para enfermedades poco frecuentes, algunas de las cuales han recibido pruebas de soporte experimental. Investigando las causas de la migraña, dicho investigador extrajo varias piezas de evidencia a partir de títulos de artículos presentes en la literatura biomédica.

Algunas de esas claves fueron:

- El estrés está asociado con la migraña.
- El estrés puede conducir a la pérdida de magnesio.
- Los bloqueadores de canales de calcio previenen algunas migrañas.
- El magnesio es un bloqueador natural del canal de calcio.
- La depresión cortical diseminada (DCD) está implicada en algunas migrañas.
- Los niveles altos de magnesio inhiben la DCD.
- Los pacientes con migraña tienen una alta agregación plaquetaria. El
- magnesio puede suprimir la agregación plaquetaria.

Estas claves sugieren que la deficiencia de magnesio podría representar un papel en algunos tipos de migraña, una hipótesis que no existía en la literatura y que Swanson encontró mediante esas ligas. De acuerdo con Swanson, estudios posteriores han probado experimentalmente esta hipótesis obtenida por text mining con buenos resultados.

10.6 Internet: Web Mining

Una de las aplicaciones de la minería de datos consiste en aplicar sus técnicas a documentos y servicios Web, lo que se denomina comúnmente con el término inglés web mining (minería de Web). Cada vez que un usuario visita un sitio Web va dejando todo tipo de “huellas” Web (direcciones de IP, navegador, galletas, etc.) que los servidores automáticamente almacenan en una base de datos (log). Las herramientas de web mining analizan y procesan esta abundante cantidad de datos para producir información significativa, por ejemplo, cómo es la navegación de un cliente antes de hacer una compra en línea. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, metadatos o hiperligas, investigaciones recientes usan el término multimedia data mining (minería de datos multimedia) como una instancia del web mining para tratar ese tipo de datos.

Los accesos totales por dominio, horarios de accesos más frecuentes y visitas por día, entre otros datos, son registrados por herramientas estadísticas que complementan todo el proceso de análisis del web mining.

También es muy importante como los link en los sitios Web son utilizados. Se puede saber cuantos links debe pasar el usuario en una página hasta llegar al contenido deseado, así, si se encuentra que una gran cantidad de usuarios acceden a un link alejado de la página principal se puede poner un acceso directo desde la misma y así ahorrar tiempo a los navegantes y conseguir un mayor beneficio o incluir publicidad en los links que se visitarán más frecuentemente.

Normalmente, la minería de datos de Web puede clasificarse en tres dominios de extracción de conocimiento de acuerdo a la naturaleza de los datos:

1. *Web content mining* (minería de contenido web). Es el proceso que consiste en la extracción de conocimiento del contenido de documentos o sus descripciones. La localización de patrones en el texto de los documentos, el descubrimiento del recurso basado en conceptos de indexación o la tecnología basada en agentes también pueden formar parte de esta categoría.
2. *Web structure mining* (minería de estructura web). Es el proceso de relacionar el conocimiento de la organización del *www* y la estructura de sus ligas.
3. *Web usage mining* (minería de uso web). Es el proceso de extracción de modelos interesantes usando los logs de los accesos a la web.

Algunos de los resultados que podrían obtenerse tras la aplicación de los diferentes métodos de web mining a una página ficticia son:

El 85% de los clientes que acceden a /productos/home.html y a /productos/noticias.html acceden también a /productos/historias_suceso.html. Esto podría indicar que existe alguna noticia interesante de la empresa que hace que los clientes se dirijan a historias de suceso. Igualmente, este resultado permitiría detectar la noticia sobresaliente y colocarla quizá en la página principal de la empresa o también se pueden observar casos donde los clientes que hacen una compra en línea cada semana en /compra/producto1.html tienden a ser de sectores de la población determinado, como estudiantes, pensionistas, funcionarios u otros. Esto podría resultar en proponer diversas ofertas a este sector para conseguir un potenciamiento en compras por parte de estos grupos.

El sesenta por ciento de los clientes que hicieron una compra en línea en /compra/producto1.html también compraron en /compra/producto4.html después de un mes. Esto indica que se podría recomendar en la página del producto 1 comprar el producto 4 y ahorrarse el costo de envío de este producto.

Radio personalizada en Internet: Last.fm

Last.fm es una radio vía Internet y además un sistema de recomendación de música que construye perfiles y estadísticas sobre gustos musicales, basándose en los datos enviados por los usuarios registrados. En la radio se puede seleccionar las canciones según las preferencias personales (de acuerdo a un algoritmo y a las estadísticas) o de otros usuarios. El servicio es de código abierto. Se fusionó con su proyecto hermano *Audioscrobbler* en agosto de 2005.

Un usuario de Last.fm puede construir un perfil musical usando dos métodos: escuchando su colección musical personal en una aplicación de música con un plugin de Audioscrobbler, o escuchando el servicio de radio a través de Internet de Last.fm, normalmente con el reproductor de Last.fm. Las canciones escuchadas son añadidas a un registro desde donde se calcularán los gráficos de barras de tus artistas y canciones favoritos, además de las recomendaciones musicales.

Las recomendaciones son calculadas usando un algoritmo colaborativo de filtrado, así los usuarios pueden explorar una lista de artistas no listados en su propio perfil pero que si aparecen en otros usuarios con gustos similares. Last.fm también permite a los usuarios manualmente recomendar discos específicos a otros usuarios (siempre que el disco esté incluido en la base de datos). Además, Last.fm soporta etiquetaje de artistas por el usuario final. Los usuarios pueden explorar vía etiquetas, pero el beneficio más importante es la radio etiquetada, permitiendo a los usuarios escuchar música que ha sido etiquetada de una manera determinada. Este etiquetaje puede ser por género ("garage rock, pop, etc."), humor ("relajado"), característica artística ("barítono"), o cualquier otra forma de clasificación hecha por el usuario final. Quizá la característica más usada por la comunidad de Last.fm es la formación de grupos de usuarios con algo en común (por ejemplo, militancia en otro foro de Internet). Last.fm generará un perfil de grupos similar a los perfiles de los usuarios, mostrando una amalgama de datos y mostrando listas con los gustos globales del grupo.

Los sellos musicales y los artistas son ayudados a promocionarse en Last.fm, porque el filtraje y recomendación son características que hacen que la música sea escuchada por usuarios que le gusten artistas similares. El stock musical de Last.fm contiene más de 100.000 canciones.

Como un sistema masivo de puntuación musical, Last.fm tiene varias ventajas sobre las listas musicales tradicionales. Mientras las listas tradicionales miden el éxito de una canción por el número de unidades vendidas y de reproducciones de radio, Last.fm lo mide por el número de gente que reproduce la canción.

Flickr

Problemática: Flickr es un sitio web de organización de fotografías digitales y red social. El servicio es utilizado extensamente como depósito de fotos. Además, el sistema de Flickr emplea técnicas de clustering de datos para agrupar las imágenes por etiquetas o tags (al igual que Last.fm). Simplemente son palabras que permiten definir algo. Por ejemplo si subimos una imagen de la Playa Punta de Palma de Izabal, Guatemala, puedo clasificarla con las siguientes etiquetas "playa" "mar" "izabal" y "guatemala".

Resultado: Pero Flickr es más que un simple sitio Web donde poder colgar tus fotos, puedes crear un perfil de usuario y encontrar gente alrededor del mundo con gustos similares a los tuyos y agregarlos a tu lista de contactos. También almacena diariamente una colección sobre las mejores fotos que se van colgando en el servidor sin que intervenga ninguna persona. Así, se consigue que estas sean de una gran calidad según las visitas recibidas, notas de otros usuarios, rating del usuario que la colgó, etc.

IV. Conclusiones

- Se determinó que la minería de datos es el conjunto de herramientas y técnicas de análisis de datos que permiten crear escenarios, de los cuales se puede obtener información útil para la toma de decisiones a nivel gerencial.

- Las técnicas que utiliza la minería de datos para la exploración consisten en la identificación de patrones.

- El proceso de la minería de datos genera conocimiento por medio de la depuración, enriquecimiento y transformación de datos que sirve para la creación de un modelo en el que se evalúa un conjunto de casos.

V. Recomendaciones

- Promover el conocimiento de la minería de datos, ya que puede ser utilizada para minimizar costos o para incrementar las ganancias de un negocio.

- La mejor manera de sacarle provecho a la minería de datos es utilizándola en conjunto, sin ir más allá podemos utilizarlo con un data warehouse como simulación, obteniendo ventajas de las técnicas de depuración de datos.

- Para obtener un mejor resultado conviene hacer diferentes aplicaciones o casos de estudio de acuerdo a la coyuntura o realidad en que vivimos.

VI. Biografía

1. Davenport, T. H. y Prusak, L. Working knowledge: How organizations manage what they know. Boston, EUA: Harvard Business School Press, 1998.

2. Moxon, B. (1996). Defining Data Mining. DBMS Online DBMS Dara Warehouse Supplement (August 1996). Disponible en: URL : http://www.dbmsmag.com/9608d53.html

3. Sofia J. Vallejos: Minería de Datos. [Trabajo pregrado en Internet] Argentina: Universidad Nacional del Nordeste; 2006 [citada 03 jul 2015]. 33p. Disponible en: URL: http://exa.unne.edu.ar/informatica/SO/Mineria_Datos_Vallejos.pdf.

4. Calderón Méndez N. MINERÍA DE DATOS UNA HERRAMIENTA PARA LA TOMA DE DECISIONES [Tesis titulada en Internet]. Guatemala: Universidad de San Carlos de Guatemala; 2006 [citada 04 jul 2015]. 78p. Disponible en: URL: http://biblioteca.usac.edu.gt/tesis/08/08_0307_CS.pdf.

5. José A. Gallardo Arancibia. Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM) [Tesis Doctoral en Internet] España: Universidad Politecnica de Madrid; 2009 [citada 17 Jul 2015]. Disponible en: URL: http://oa.upm.es/1946/1/JOSE_ALBERTO_GALLARDO_ARANCIBIA.pdf

6. 5. Chapman P., (NCR), Clinton J., (SPSS) Kerber R., (NCR), Khabaza T. (SPSS), Reinartz T. (DaimlerChrysler), Shearer C. (SPSS), and Wirth R.(DaimlerChrysler).CRISP-DM 1.0 step-by-step data mining guide [Guia en Internet] USA y The Netherlands:SPSS;2000 [citada 07 Jul 2015] Disponible en: URL: https://www.kde.cs.uni-kassel.de/lehre/ws2015-16/kdd/files/CRISPWP-0800.pdf

7. Portal www.kdnuggets.com, “consulta sobre metodologías utilizadas en Data Mining”, [en línea] 2007 [citada 13 Jul 2015], disponible en: URL: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm [Citada 10 Jun 2008].

8. Portal www.aeipro.com, “METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE DATA MINING”, [en línea] 2010 [citada 17 Jul 2015], disponible en: URL: http://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf.

9. Portal www.technet.microsoft.com, “Soluciones de minería de datos”, [en línea] 2007 [citada 17 Jul 2015], disponible en: URL: http://technet.microsoft.com/es-es/library/ms174861.aspx

10. Guillermo G. Molero Castillo. Desarrollo de un modelo basado en técnicas de Minería de Datos para clasificar zonas climatológicamente similares en el estado de Michoacán [Tesis Magistral en Internet]. Mexico: Universidad Nacional Autónoma de Mexico; 2008 [citada 20 jul 2015]. 163p. Disponible en: URL: http://www.geologia-feflow.unam.mx/documentos/tesis%20mineria%20de%20datos.pdf

11. Hernandez J.,Ramirez M.J. y Ferri (2004).Introducción a la Minería de Datos. Pearson Educación. Editorial Pearson Hall, pp.680,ISBN:84-205-4091-9, Madrid, España.

12. Vazirgiannis M., Halkini M. y Gunopulos D. (2003). Uncertainty Handling and Quality Assessment in Data Mining. Advanced Information and Knowledge Processing, editado por Springer-Verlag, pp. 226, ISBN: 1-85233-655-2, Heidelberg, Alemania.

13. Sumathi S. y Sivanandam S. (2006). Introduction to Data Mining and its App.lications. Studies in Computational Intelligence, 29, editado por Springer-Verlag, pp. 828, ISBN: 3-540-34350-4, Heidelberg, Alemania.

14. Berry M. y Linoff G. (2004). Data Mining Techniques: for marketing, sales, and customer relationship management. 2da edición por Wiley Publishing, Inc.,pp. 643, ISBN: 0-471-47064-3, Indiana, Estados Unidos.

15. Larose D. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc., pp. 222, ISBN: 0-471-66657-2, New Jersey, Estados Unidos.

16. Jain A., Murty M. y Flynn P. (1999). Data Clustering: A Review. ACM Computing Surveys, 31, 3, 264-323, ISBN: 0360-0300, New York, Estados Unidos.

17. Witten I. y Frank E. (2005). Data Mining: Practical machine learning tools and techniques. 2da ed. por Morgan Kaufmann Series in Data Management Systems, pp. 525, ISBN: 0-12-088407-0, Estados Unidos.

18. Hand D., Mannila H. y Smyth P. (2001). Principles of Data Mining, editado por The Massachusetts Institute of Technology Press., pp. 546, ISBN: 0-262-08290-x, Massachusetts, Estados Unidos.

19. Juan J. Montaño Moreno. Redes Neuronales Artificiales Aplicada al Analisis de Datos [Tesis Doctoral en Internet] PALMA DE MALLORCA: UNIVERSITAT DE LES ILLES BALEARS; 2002 [citada 15 Jul 2015]. Disponible en: URL: http://www.tesisenred.net/bitstream/handle/10803/9441/tjjmm1de1.pdf?sequence=1

20. Fodor, J.A. y Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: Acritical analysis.pp. 28, ISBN: 0-262-66064-4, Estados Unidos.

21. Portal www.medintensiva.org, "Redes neuronales artificiales en Medicina Intensiva. Ejemplo de aplicación con las variables del MPM II", [en línea] 2005 [citada 17 Jul 2015], disponible en: URL: http://www.medintensiva.org/es/redes-neuronales-artificiales-medicina-intensiva-/articulo/13071859/

22. Portal www.personales.upv.es, "ALGUNAS APLICACIONES DE REDES NEURONALES ARTIFICIALES EN DOCUMENTACIÓN", [en línea] 2014 [citada 17 Jul 2015], disponible en: URL: http://personales.upv.es/ccarrasc/doc/2013-2014/Redes%20Neuronales%20Excalibur%20-%20Nativ.%20Noverges/redes.htm#1.1

23. Damián J. Matich.Redes Neuronales: Conceptos Básicos y Aplicaciones [Cátedra en Internet]. Argentina:Universidad Tecnológica Nacional; 2001 [citada 18 jul 2015]. 55p. Disponible en: URL: https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf

24. Portal www.webmining.cl, "200+ herramientas gratuitas de Estadística y Data Mining", [en línea] 2011 [citada 19 Jul 2015], disponible en: URL: http://www.webmining.cl/2011/03/200-herramientas-gratuitas-de-estadistica-y-data-mining/

25. Portal www.culturacrm.com, "Cinco herramientas de Data Mining", [en línea] 2016 [citada 19 Jul 2015], disponible en: URL: http://culturacrm.com/data-mining/cinco-herramientas-data-mining/

26. Portal www.mprende.co, "6 herramientas gratuitas para análisis de datos", [en línea] 2015 [citada 19 Jul 2015], disponible en: URL: http://mprende.co/gesti%C3%B3n/6-herramientas-gratuitas-para-an%C3%A1lisis-de-datos

27. Portal www.perustat.com, “El uso de R se dispara según la encuesta de Rexer Analytics”, [en línea] 2013 [citada 21 Jul 2015], disponible en: URL: http://perustat.com/blog/el-uso-de-r-se-dispara-segun-la-encuesta-de-rexer-analytics/

28. Portal www.r-bloggers.com, “R tops KDNuggets data analysis software poll for 4th consecutive year”, [en línea] 2014 [citada 21 Jul 2015], disponible en: URL: https://www.r-bloggers.com/r-tops-kdnuggets-data-analysis-software-poll-for-4th-consecutive-year/

29. Portal www.kdnuggets.com, “Languages for analytics / data mining / data science”, [en línea] 2014 [citada 21 Jul 2015], disponible en: URL: http://www.kdnuggets.com/polls/2014/languages-analytics-data-mining-data-science.html

30. Portal www.kdnuggets.com, “Four main languages for Analytics, Data Mining, Data Science”, [en línea] 2014 [citada 20 Jul 2015], disponible en: URL: http://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html

31. Portal www.kdnuggets.com, “Analytics, Data Mining, Data Science software/tools used in the past 12 months”, [en línea] 2015 [citada 20 Jul 2015], disponible en: URL: http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html

32. Portal www.kdnuggets.com, “Analytics, Data Mining, Data Science software/tools used in the past 12 months”, [en línea] 2015 [citada 21 Jul 2015], disponible en: URL: http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html

33. Tom Khabaza. Área de Cascos Duros: Mitos y Trampas de la Minería de Datos [en línea] SPS: Director de la Minería de Datos; 2010 [citada 22 Jul 2015]. Disponible en: URL: http://dunyateknolojisi.com/Uploads/alpoges/spss_en_espagnol.pdf