INTRODUCCION A LA INTELIGENCIA DE NEGOCIOS CON ENFASIS EN UNA SELECCION DE ALGORITMOS DE MINERIA DE DATOS

JHOSEP DAVID PARRA MORENO
JONATHAN ANTONIO YARA PEÑALOZA

UNIVERSIDAD TECNOLOGICA DE PEREIRA
FACULTAD DE INGENIERIAS
INGENIERIA DE SISTEMAS Y CIENCIAS DE LA COMPUTACIÓN
PEREIRA
2014

INTRODUCCION A LA INTELIGENCIA DE NEGOCIOS CON ENFASIS EN UNA SELECCION DE ALGORITMOS DE MINERIA DE DATOS

JHOSEP DAVID PARRA MORENO
JONATHAN ANTONIO YARA PEÑALOZA

CARLOS AUGUSTO MENESES ESCOBAR
PROFESOR

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERIAS
INGENIERÍA DE SISTEMAS Y CIENCIAS DE LA COMPUTACIÓN
PEREIRA
2014

Tabla de contenido

Tabla de ilustraciones

CAPITULO I

1. Titulo

Introducción a la Inteligencia de Negocios con énfasis en una selección de algoritmos de Minería de Datos

2. Definición del problema

La información, como componente/activo de las organizaciones, se ha convertido con el paso de los años en uno de los recursos más valiosos. Cada transacción, movimiento, o proceso debe estar soportado en documentación, datos o cifras. Y tal como ha crecido la trascendencia de la información así mismo ha crecido el volumen de ésta. Cantidades enormes de datos, que son la esencia de la empresa misma, se mantienen guardados en distintas ubicaciones físicas, en ocasiones aislados entre sí. Lo que interesa a las organizaciones es poder unificar la información y obtener extractos de ésta que puedan ser aprovechados para tomar las decisiones que las lleven por el camino correcto.

Actualmente la mayoría de empresas posee diversos mecanismos de captura de datos. Sin embargo, estos datos no son bien administrados en la mayoría de ocasiones. Generalmente son usados o alcanzados por muy pocas personas dentro de la compañía, las cuales además, no tienen relación entre sí. Si se quiere utilizar los datos de la empresa para generar un valor agregado a la misma, lo primero que se debe tener en cuenta es que los datos de un área por sí sola no traerán información relevante para la compañía. Los datos deben ser compartidos entre las diferentes

áreas, con lo cual se generaría información más valiosa y más relevante a la disposición de toda la compañía.

Hasta hace algún tiempo, las empresas que lograban alcanzar este estado de manejo de información se veían como las más fuertes ante sus similares. Sin embargo hoy en día no es suficiente con poseer la información de la empresa. Ahora, además de capturar los datos y transformarlos en información relevante, se necesita un nuevo nivel en el manejo de la información. El análisis y la toma de decisiones.

El día a día de una empresa se vive en la toma de decisiones. Y no sólo las decisiones que toman el área gerencial o la cúpula administrativa. Todos los días un gran porcentaje de empleados de la compañía se ve en la obligación de tomar una decisión en su puesto de trabajo. Independiente de que esta decisión traiga una repercusión muy grande o mínima, siempre se debe tomar la mejor opción, dejando lo más cercano a cero el margen de error. ¿Pero en qué se soportan los empleados para la toma de decisiones? En la mayoría de los casos, estas personas se apoyan en datos planos obtenidos de una aplicación (Hojas de cálculo, archivos de texto plano), en informes de sus superiores, o en base a su experiencia (Cuando el tiempo de respuesta es mínimo). Tales suministros de información en ocasiones no cuentan con datos actualizados y/o disponibles a lo largo del tiempo.

Ante esta problemática, surgen dos alternativas que en conjunto pueden llegar a ofrecer más que una mera solución. La inteligencia de negocios (BI, por sus siglas en inglés), como herramienta gerencial; y la minería de datos (DM, por sus siglas en inglés), como soporte técnico a la BI.

La inteligencia de negocios ofrece a las organizaciones la habilidad de transformar los datos en información, y la información en conocimiento, para que la toma de decisiones se haga de la manera más óptima. Hablando más específica y técnicamente, se refiere a un conjunto de metodologías, aplicaciones y tecnologías para reunir, filtrar y transformar datos de los sistemas transaccionales e información desestructurada (Los

suministros de información como los archivos de texto, hojas de cálculo, etc) en información con estructura definida (Cubos OLAP, data warehouses, data marts, entre otro), para que se pueda analizar y dar soporte a la toma de decisiones sobre el negocio.

La minería de datos se encarga de extraer la información de los procesos, y además de encontrar relaciones, comportamientos y patrones, invisibles a simple vista, dentro de las fuentes de datos. [1] El marco de trabajo o metodología más usada en la actualidad referente a minería de datos se conoce como CRISP-DM. Todo un set de fases que indican en distintos niveles de abstracción qué se debe realizar en un proyecto de minería de datos.

Por lo tanto, ¿Será posible lograr tener un documento de referencia para la toma de decisiones a nivel empresarial mediante el uso de inteligencia de negocios como herramienta gerencial y la minería de datos como complemento técnico?

3. Justificación

La inteligencia de negocios es una herramienta de gran potencial, la cual está pasando por un momento de gran aceptación e inclusión en grandes compañías. A pesar de que muchas organizaciones implementan o quieren implementar herramientas de inteligencia de negocios, no hay muchas personas que conozcan a fondo el tema, o que puedan implementar dichas herramientas. Además, la incursión en el tema para nuevas personas es un poco complicada, ya que la documentación se encuentra dispersa y sin relación entre sí.

Tal es el posicionamiento de la inteligencia de negocios en las grandes compañías, que cada día se necesitan más profesionales que puedan ayudar a establecer, soportar o implementar esta herramienta dentro de las organizaciones. Pero a pesar de la gran demanda que presenta este campo, muy pocas personas en el interior de las

organizaciones poseen los conocimientos básicos sobre el tema. Los nuevos profesionales en campos de las ciencias de la computación terminan sus estudios de pregrado sin tener idea alguna de este campo de acción.

Referente a minería de datos, al momento de elegir un marco de trabajo para desarrollar proyectos de este tipo la respuesta es básicamente la misma, la metodología CRISP-DM. De este tema se encuentra literatura amplia debido a la gran aceptación que tiene en los grupos especializados de minería de datos, como se puede constatar en la comunidad virtual KD Nuggets [3] y en una serie de encuestas que realizaron entre el 2002 y el 2007, en donde se evidencia el crecimiento de la población que prefería CRISP-DM por sobre sus demás rivales.

Además no solamente es alta la demanda de procesos de inteligencia de negocios y subsecuentemente de minería de datos, sino también la oferta de técnicas y algoritmos de minería de datos y su diversidad de uso. Debido a esto, es importante saber delimitar el alcance de una determinada técnica y tener conocimiento de sus características para así llegar con la certeza de que se obtendrán los resultados esperados a través del camino más óptimo.

Tomando en cuenta estas consideraciones, se pretende tener una guía de estudio o de partida inicial, para las personas que quieran ingresar al campo de la inteligencia de negocios y la minería de datos, en la cual se podrá ver de forma centralizada y ordenada la información esencial acerca de ambos temas. ***(Se eliminaría)***

Aquella persona que lea este documento encontrará una convergencia de términos relacionados con la toma de decisiones a distintos niveles conceptuales (Desde lo gerencial hasta lo técnico). Quien lea este documento tendrá la ventaja teórica de no tener que recurrir a fuentes externas para tener conocimiento sobre cuatro temas que se consideran pilares: ***Sistemas de Soporte a la Decisión, Inteligencia de Negocios, Minería de Datos, y Almacenamiento de Datos***.

Algunos temas simplemente se mencionan, o se tratan de manera superficial, pero se deja como constancia la referencia de la fuente de la cual proviene el contenido, para así indagar más profundamente, pero sólo si se requiere un conocimiento más profundo.

4. Objetivos

4.1. Objetivo general

Elaborar un documento que permita referenciar a la inteligencia de negocios y la minería de datos como herramientas ideales en el proceso de toma de decisiones y descubrimiento de información en grandes volúmenes de datos.

4.2. Objetivos específicos

- Presentar un contexto histórico del desarrollo de la inteligencia de negocios a través de una inducción a los DSS (Sistemas de soporte a la decisión) y su evolución.
- Describir la definición, características, ventajas e implicaciones de las herramientas de inteligencia de negocios en una organización.
- Describir la definición, características, ventajas e implicaciones de la minería de datos.
- Describir detalladamente las fases de la metodología CRISP-DM como marco de trabajo para proyectos de minería de datos.
- Listar un conjunto de los algoritmos de minería de datos más comunes y en qué casos son útiles.
- Elaborar el documento planteado en el objetivo general.

CAPITULO II

5. Marco referencial

5.1. Marco de antecedentes

En la década de los sesenta, cuando las grandes organizaciones empezaban a computarizar algunos procesos de sus empresas, se empezaron a construir diversos sistemas de información de administración (Management Information Systems, MIS), cuyo objetivo era poner a disposición la información obtenida para el proceso de toma de decisiones por parte de la cúpula administrativa. El problema radicaba en la brecha de conceptos que existía entre desarrolladores y administradores: Los primeros realizaban sistemas grandes e inflexibles, y los reportes de salida solían ser extensos y con poca información útil para los segundos.

Tras varios años de acuñar distintas definiciones, aparece a principios de los años 70 el término “sistema de soporte a la decisión” (Decision Support System, DSS), en manos de Gorry y Scott Morton. La primera definición formal que hicieron fue que DSS era un “Sistema que soporta a los administradores en situaciones donde hay que tomar decisiones no estructuradas”. [4]

El factor clave para la trascendencia que adquirieron los DSS fue el soporte que daban a los administradores y gerentes al momento de tomar decisiones sin estructura fija, decisiones sin un plan concreto, cuya resolución no siguiera una secuencia de pasos lógicos (Algoritmo).

Cabe puntualizar que la definición de DSS ha sido tan amplia desde sus inicios, que ha sido preferible estudiarlo a través de sus subtipos y campos subyacentes: Sistemas de

Administración de Base de Datos (Data Base Management Systems, DBMS), Interfaces Gráficas de Usuario (Graphic User Interface, GUI), e incluso Sistemas de Información Ejecutiva/Inteligencia de Negocios (Executive Information System/Business Intelligence, EIS/BI) y Almacenamiento de Datos (Data Warehousing, DW).

Debido al alcance que tiene este documento no se va a entrar en detalles de la evolución que tuvo el campo de DSS en los años siguientes, durante los últimos años de los 70 y los 80, entre los que podemos destacar algunas mejoras en las GUI; avances en los DBMS (Surgimiento de los modelos relacionales, entidad relación, entre otros); modelado lógico a través de predicados, etc.

Gracias a los avances tecnológicos en las décadas de los 70 y los 80, surgió el predecesor directo de la inteligencia de negocios, los sistemas de información ejecutiva (EIS). Enfocados a todo el nivel gerencial de una compañía, los EIS permitían al gerente, a partir de informes jerárquicos, descubrir las fallas de un área crítica de la compañía. Además, innovó con el concepto de la multidimensionalidad de los datos, también denominada como 'Cubo de Datos', lo que más tarde pasaría a ser conocido como el cubo OLAP (On-Line Analytical Processing).

La popularidad de los EIS creció y se sostuvo hasta la década de los 90, época en la cual, cualquier empresa (con la capacidad de hacerlo) contaba con un EIS en su abanico de herramientas de tecnología. Sin embargo, debido a algunos cambios necesarios en la estructura de los EIS se empezó a hablar de un nuevo término, la Inteligencia de negocios (BI).

Por otro lado, algunos ejemplos iniciales de lo que se conoce hoy día como minería de datos se dieron incluso siglos antes del desarrollo de la computación. El término actual proviene de un trabajo establecido a finales de los 80 por Gregory Piatetsky-Shapiro en un taller sobre Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases, KDD), que hacia 1995 pasó a llamarse Conferencia Anual sobre Descubrimiento del Conocimiento y Minería de Datos. [19] [7]

Siendo un campo de estudio tan relativamente joven, los proyectos que incluyen minería de datos han sufrido los mismos atascos que en un principio sufrieron los proyectos de ingeniería de software: Retrasos en los proyectos, baja eficiencia y la falla casi permanente en la comunicación con el cliente, que no permite satisfacer las necesidades del usuario. Para la ingeniería del software la situación mejoró considerablemente con la definición de nuevas metodologías. Para la minería de datos el caso es similar. [5]

La entrada del nuevo milenio trajo consigo para este campo un hito. Con el escenario previamente establecido, surgía CRISP-DM (Cross Industry Standard Process for Data Mining), una metodología o modelo de procesos que en la teoría se pensaba independiente de la industria, la herramienta y la aplicación. Para su creación, se conformó el Grupo de Interés Especial (Special Interest Group, SIG), apoyado por multinacionales privadas (Compañías como DaimlerChrysler o SPSS) y por la Comisión Europea. [2]

Desde sus inicios triunfó dentro de la comunidad de la minería de datos, siendo considerado en la actualidad el estándar de facto. Distintas encuestas extraoficiales realizadas por la comunidad virtual KD Nuggets en el 2002 [8], el 2004 [9], y el 2007 [10] así lo confirman. Probablemente uno de los criterios que soporta su éxito es que está basado en 5 A's, SEMMA y Two Crows, tres de las metodologías anteriores más usadas.

5.2. Marco teórico

5.2.1. Sistemas de soporte a la decisión

La definición de sistemas de soporte a la decisión o DSS abarca una generación de herramientas gerenciales tan amplia que dar una definición formal que los abarque a todos resulta insatisfactoria e incompleta.

La clasificación actual de los DSS ha ido evolucionando desde sus inicios y cada tipo diferente de DSS ha tenido su momento. La primera clasificación de DSS fue realizada en 1975 por el estudiante de doctorado del MIT Steven L. Alter, a partir de una muestra de 56 DSS agrupados en 7 categorías que diferían entre sí por sus operaciones genéricas.

Alter cita en su trabajo que a pesar de que por definición un DSS está pensado para soportar la toma de decisiones, "...No es una categoría homogénea. Muy por el contrario, muchos de los sistemas en la muestra diferían vastamente en lo que hacían y en cómo lo hacían. Los temas clave a la hora de la implementación varían a través de diferentes tipos de DSS…"

La primera clasificación de Alter incluía las siguientes categorías: **Sistemas de cajón de archivos**, que proveían acceso a objetos de datos; **Sistemas de análisis de datos**, que apoyaban la manipulación de datos con herramientas computacionales adaptadas a una tarea específica; **Sistemas de información de análisis**, que proveían acceso a bases de datos orientadas a la decisión; **DSS contables y financieros basados en modelo**, que calculaban las consecuencias de las acciones posibles; **DSS representacionales basados en modelo**, que estimaban las consecuencias de acciones en base a modelos de simulación; **DSS de optimización basados en modelo**, que proveían una solución óptima y además poseían restricciones que dirigían el proceso de toma de decisión; y finalmente **DSS de sugerencia basados en modelos lógicos**, que realizaban procesamiento lógico tomando como materia prima una tarea bien estructurada y bien entendida.

5.2.2. Tipos de DSS [6] [11]

Arnott y Pervan, en su artículo "A critical analysis of Decision Support Systems research" hacen una clasificación de acuerdo a la evolución histórica de los DSS de la cual salen los siguientes tipos:

- DSS Personales (Personal DSS, PDSS)
- DSS Grupales (Group DSS, GDSS)
- DSS a las Negociaciones (Negotiation DSS, NDSS)
- DSS Inteligentes (Intelligent DSS, IDSS)
- Sistemas de Información Ejecutiva e Inteligencia de Negocios (Executive Information Systems and Business Intelligence, EIS/BI)
- Bodegas de Datos (Data Warehouses, DW)
- DSS Basados en Gestión del Conocimiento (Knowledge Management-based DSS, KM DSS)

De la anterior taxonomía, los dos primeros han dominado los estudios de investigación durante los últimos años. Aun así, resulta curioso que las organizaciones apuesten más hacia los EIS/BI y los DW. Esto es evidencia de la gran brecha que existe en el campo de DSS entre la academia y la industria.

5.2.3. EIS/BI.

6.2.3.1. Sistemas de Información Ejecutiva EIS

Los EIS son sistemas de soporte a la decisión orientados a los datos, desarrollados para usuarios de perfil administrativo dentro de una organización, cuya principal función es brindar información relevante sobre un área específica de la organización a sus usuarios. El acceso a sistemas de archivos simples a través de consultas y herramientas de recuperación de datos hacen parte del nivel de funcionalidad más bajo o elemental. Los sistemas de almacenamiento que permiten la manipulación de datos a través de herramientas computacionales adaptadas a tareas específicas comprenden otra parte de la funcionalidad media. El nivel más alto viene dado por el procesamiento analítico en línea (On-Line Analytical Processing, OLAP). Los sistemas EIS impulsaron el desarrollo de los DW.

Los sistemas de información ejecutiva o EIS (por sus siglas en inglés) son el padre de la inteligencia de negocios. Después de un largo proceso de crecimiento de las herramientas de toma decisiones, se llega a lo que parece ser el suministro completo de información de una compañía, una herramienta con la cual puede contar desde el gerente general de la empresa hasta un auxiliar de cualquier área de la misma.

Desde que se empezó a hablar de la inteligencia de negocios como herramienta tecnológica útil en la toma de decisiones de una compañía, se establecieron como prioridades en el diseño la interacción del usuario final con la herramienta y el buen manejo de los requerimientos hechos por ellos. De este modo surgieron consultas más manejables, una interfaz gráfica más agradable y una información más acorde a la que es solicitada por el cliente.

Teniendo definidas las prioridades de las herramientas de BI se propuso como objetivo claro de su implementación, brindar ayuda para tomar mejores y más rápidas decisiones en todos los niveles de la empresa, lo cual es posible gracias a la optimización del tiempo que hay entre la recolección de la información y la disponibilidad de la misma.

5.2.3.2. Inteligencia de Negocios BI

Desde la implementación de los sistemas de soporte a la decisión, se ha hecho más fuerte la idea de que la información también es un recurso de la empresa, y que ésta, debe ser compartida y utilizada por la mayor cantidad de personas dentro de la organización. La información no es un recurso de alguien en particular dentro de la organización. Por el contrario, los recursos ricos en característica intelectual aumentan su valor proporcionalmente a la frecuencia de su uso.

Es aquí donde BI juega un gran papel en el ámbito comercial de las empresas. La evolución de la información dentro de las compañías ha crecido, al igual que la

tecnología, exponencialmente y lo que hace 10 años era una gran ventaja desde el punto de vista competitivo de las organizaciones, hoy es un simple primer paso de lo que se necesita para tener un valor agregado en la competencia del mercado.

Cuando se extraen datos de cualquier tipo de suministro de información con el que cuente la empresa, nos encontramos con datos planos, dispersos y segmentados, los cuales por sí solos no brindan una información útil del entorno del negocio. Además, estos datos provienen de distintas fuentes, como archivos planos, hojas de cálculo, archivos XML, bases de datos, entre otras. Sin embargo, así estos datos se vean descartables o sin utilidad, son el primer paso para la implementación de una Herramienta BI. Los datos planos son el primer suministro de información de mi sistema, en la arquitectura de una herramienta BI identificar estas fuentes de datos es el inicio para el proceso de implementación de la herramienta.

Una vez se hayan identificado los diferentes suministros de información se debe tomar la información de cada uno de ellos, pero debido a su dispersión e irregularidad, esta tarea no es tan sencilla como suena. Para poder capturar está información y hacerla útil para la empresa, se debe implementar una herramienta ETL (Por sus siglas en inglés, Extract, Transform and Load), cuya función es capturar los datos de las diferentes fuentes; organizarlos, limpiarlos, depurarlos, transformarlos; y finalmente cargarlos a un almacén o bodega de datos.

Después de tener nuestros datos depurados, transformados y cargados en la bodega de datos, ya cuento con un repositorio de datos ordenados, coherentes y de igual estructura en mi bodega. Estos datos se ven representados visualmente por cubos de datos, modelos multidimensionales, tablas, entre otros.

Por último en la arquitectura de una herramienta de BI se encuentran las herramientas de análisis y reportes. En esta parte es donde entra a interactuar el usuario final con la información generada en todo el proceso, creando reportes personalizados, cuadros de mando o gráficos, realizando consultas y análisis sobre la información. Estas

herramientas brindan la información necesaria a los encargados de los procesos para generar estrategias comerciales y toma de decisiones a partir del conocimiento credo por la misma empresa.

Ya no es grande y poderoso el que posee la información, si no el que la analiza y saca mejor partido de ella. Ésta es la propuesta que nos brindan las herramientas de BI, utilizar el conocimiento generado por nuestros propios datos en pro de la organización, siendo usada como una ventaja competitiva a la hora de tomar decisiones.

5.2.3.3. El valor de la inteligencia de negocios

Con la implementación de una solución de BI, la compañía se ve beneficiada en diferentes aspectos. Siendo la parte comercial la más beneficiada a simple vista. El valor agregado de una herramienta BI se puede ver en el ahorro de los costos normales de la empresa. Al contar con un suministro constante y 'en línea' de reportes, se reduce considerablemente el tiempo de generación de los mismos, los cuales son de vital importancia a la hora de la toma de decisiones, por básicas que sean. Además, como cada usuario, independiente de su cargo o nivel jerárquico en la compañía, cuenta con el acceso a los reportes necesarios para el completo desarrollo de sus actividades; una solución BI provee eficiencia operacional, reduciendo tiempos de entrega de información a través del autoservicio de la misma.

A través de la minería de datos, como una de las herramientas en las que se apoya la inteligencia de negocios, se pueden mejorar las estrategias de mercadeo. Buscando tendencias en las compras, los clientes más rentables, las ofertas más aceptadas, segmentación de ventas y todos los demás reportes que se le ocurran al departamento comercial de la compañía. Al pasar menos tiempo en la elaboración de reportes, búsqueda de datos o medición de indicadores, se mejora drásticamente la eficiencia operacional, con lo cual se incrementan los ingresos de la compañía.

Tal vez uno de los valores agregados más valiosos que deja la implementación de una solución BI no se puede observar a simple vista. La mejora en la comunicación interna es sin duda uno de los frutos más valiosos que deja una herramienta BI, ya que una de las creencias de las herramientas BI es que la información de la empresa le pertenece a cada uno de los individuos, áreas y gerencias que la conforman. A través de un mismo suministro de información y estimulación de relaciones entre departamentos, las herramientas BI promueven el trabajo en equipo y la información colectiva. Variables con las cuales se pretende aumentar la productividad, eficiencia y el cumplimiento de objetivos estratégicos.

5.2.4. Data Warehousing y Data Warehouses [15]

Antes de iniciar hay que hacer un especial énfasis en las diferencias que existen entre los conceptos Data Warehousing (Que podría traducirse al español como Almacenamiento de Datos) y Date Warehouse (Almacén o Bodega de datos). Data warehousing abarca toda una arquitectura y procesos, no se limita a tener una bodega de datos. Comprende la transformación de simples datos a información útil para revelar las operaciones y rendimiento de una organización. [12]

Cada vez que se obtienen datos a partir de los sistemas operacionales para realizar reportes y análisis, se están ejecutando procesos de data warehousing. Este conjunto de procesos es lo que se encuentra tras bastidores de los grafos y pivotes que hay en BI. De hecho, BI puede ser visto como la capa de presentación de la arquitectura de data warehousing.

El error de definición recae al asociar data warehousing con una bodega de datos en vez de hacerlo con toda la arquitectura y sus procesos. Cuando se reduce el enfoque a una sola base de datos, se pierde el contexto entero de las distintas etapas de los datos. La calidad, la consistencia y la integridad de los datos sólo se logra cuando se considera una arquitectura de data warehousing entera.

Otras literaturas apuntan a que el almacenamiento de datos como tal comprende un conjunto de tecnologías (Fuentes de datos; herramientas de transformación, análisis y acceso de datos) cuyo objetivo es facilitar y agilizar el proceso de toma de decisiones para los Decision Makers. Parte de su importancia radica en que las organizaciones pueden contar con un punto central de donde pueden extraer información actualizada y ordenada históricamente.

Las bodegas de datos son definidas por Bill Inmon como “Una colección de datos orientada a un tema específico, integrada, variable en el tiempo y no volátil para ayudar al proceso de toma de decisiones gerenciales”. Inmon es considerado en el entorno de las TI como el padre de los almacenes de datos. [13]

Cabe resaltar que las características principales de los almacenes de datos (Orientación a temas específicos, integración, variabilidad con el tiempo, no volatilidad) le permiten mantener una brecha grande con respecto a sistemas similares, como lo son los DBMS, no solamente en su definición teórica, sino en la definición del diseño o al momento del modelado. [16]

Por el tamaño que alcanza una bodega de datos, es común que se divida ésta en pequeñas bodegas conocidas como Data Marts. Un Data Mart es una bodega parcial enfocada a un área o departamento de la organización y está diseñado para responder preguntas específicas de un grupo específico de usuarios. [14]

Debido a la delimitación de los datos que obliga a que sólo se incluya la información relevante para la toma de decisiones dentro de la empresa; la capacidad de aprender de los datos del pasado y predecir futuras situaciones; y el considerable bajo tiempo de respuesta en las consultas a su sistema, **se considera que la implantación de una arquitectura de Data Warehousing es el primer paso para poner en marcha una solución de tipo Business Intelligence dentro de cualquier empresa.**

5.2.4.1. Procesos ETL. [17]

Después de que se carga por primera vez la información en una bodega de datos, esta se empieza a alimentar a sí misma. Las bodegas de datos se rigen bajo un proceso de construcción denominado ETL (Por sus siglas en inglés, Extract, Transform and Load). El primer paso de esta construcción es extraer información de su fuente principal; después esta información es analizada, relacionada y transformada; y por último, el resultado de este análisis es retroalimentado al almacén de datos.

La extracción implica la recolección de datos desde las diversas fuentes que alimentan la bodega de datos. Cuando se han extraído los datos, se requiere transformarlos, proceso en el cual se limpian, filtran, depuran, manipulan y homogenizan los datos con el propósito de que compartan un formato consistente. Este proceso es necesario, ya que en ocasiones las bases de datos operacionales difieren en formato, dándose casos en los que un mismo dato llega de distintas fuentes con un nombre distinto. Al finalizar la transformación, se realiza el cargue de los datos a la bodega.

A la hora de seleccionar una herramienta que ejecute un proceso ETL de manera sencilla, es importante tener en cuenta las siguientes características:

- Manejo de grandes volúmenes de datos, importante a la hora de trasladar datos de un punto a otro en el menor tiempo posible.
- Detección de cambios en transacciones en tiempo real y sincronización de datos.
- Manejo de una variedad de datos incluyendo texto y datos sin estructura, como imágenes.
- Distribución del procesamiento en múltiples procesadores y opciones concurrentes.

5.2.4.2. OLAP [18]

Una vez la bodega de datos contiene información, es vital contar con mecanismos de acceso al alto volumen de datos que se espera que esté allí. El Procesamiento Analítico en Línea (On-Line Analytical Processing, OLAP) es un sistema que provee acceso ágil a volúmenes considerables de datos, haciendo que las consultas se hagan de manera más sencilla. Basan su funcionamiento en los cubos OLAP, estructuras de datos similares a las tablas, pero con un mayor número de dimensiones.

Son especialmente útiles a la hora de usar sentencias SQL del tipo SELECT, en contraposición a los sistemas de Procesamiento Transaccional en Línea (On-Line Transactional Proccesing, OLTP), de mayor uso con sentencias del tipo INSERT, UPDATE y DELETE. Una de sus principales características es el permitir divisar cualquier correlación existente en una gran cantidad de datos.

Cuentan con cuatro operaciones básicas para el análisis de datos:

- Drill down (Excavar, hacia adentro): Cuando se navega de un nivel jerárquico hacia uno de mayor detalle.
- Drill up (Excavar, hacia afuera): Contrario a lo anterior, se puede ir de un nivel jerárquico hacia otro superior, para ver información agregada, acumulados, entre otros.
- Slice (Corte): El corte es una columna de datos correspondiente a un valor único para uno o más miembros de una dimensión. Se puede ver como un filtro especializado para visualizar un valor particular en una dimensión.
- Dice (Dado): Mientras que la operación slice se usa con un atributo, la operación dice se puede ver más como un acercamiento (Zoom in) que selecciona todas las dimensiones, asignando valores específicos para estos. Es decir, es extraer un cubo a menor escala.

OLAP le permite a un usuario extraer y visualizar datos fácil y selectivamente desde

diferentes puntos de vista. Para facilitar el análisis, los datos OLAP se guardan una base de datos “multidimensional”. Mientras que una base de datos relacional considera cada atributo (Como el producto, las ventas regionales, y el período de tiempo) como una “dimensión” separada.

El software OLAP puede ubicar la intersección de las dimensiones (Todos los productos de la región oriental con cierto precio durante un cierto período de tiempo) y desplegarlos. Los atributos tales como los períodos de tiempo se pueden desmenuzar en subatributos.

Al usar Conectividad a Base de Datos Abierta (ODBC), los datos se pueden importar desde bases de datos relacionales existentes para crear una base de datos multidimensional para OLAP. Los productos de OLAP están diseñados usualmente para ambientes multiusuarios, con el costo basado en el número de usuarios.

5.2.5. Minería de Datos

Desde un punto de vista estrictamente académico, la minería de datos hace parte de un proceso mayor conocido como extracción de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD), siendo solamente uno de sus pasos. Usama Fayyad y otros, en un artículo de 1996 dicen que “Históricamente, la noción de hallar patrones útiles en los datos ha tenido una variedad de nombres, incluyendo, minería de datos, extracción del conocimiento, descubrimiento de información, cultivo de información, arqueología de datos, y procesamiento de patrones de datos… Desde nuestro punto de vista, KDD se refiere al proceso total de descubrimiento de conocimiento útil a partir de los datos, y la minería de datos se refiere a un paso particular del proceso.” [19]

Más adelante, en el mismo artículo, se puede observar lo siguiente: “La minería de datos es la aplicación de algoritmos específicos para extracción de patrones a partir de los datos. Los pasos adicionales en el proceso KDD, tales como la preparación de los

datos, selección de los datos, limpieza de datos, incorporación de conocimiento anterior apropiado, y la interpretación adecuada de los resultados de minería, son esenciales para asegurar que lo que se va a derivar de los datos es conocimiento útil. La aplicación “ciega” de métodos de minería de datos puede ser una actividad peligrosa, que fácilmente puede llevar al descubrimiento de patrones inválidos y sin sentido.” Se aprecia entonces una diferenciación entre los conceptos de KDD y DM a partir de establecer la importancia que tienen los otros pasos en los procesos de KDD, es decir, sin delegar una responsabilidad indebida a la minería de datos.

Teniendo en cuenta el alcance del presente documento, cuando se haga referencia a Minería de Datos se tomará como el proceso completo de Descubrimiento de Conocimiento en Bases de datos. Además, se hace un énfasis especial en dos temas propios de la minería de datos: La metodología de mayor reconocimiento en la comunidad, CRISP-DM; y un conjunto de algoritmos de importante utilidad a la hora de implementar proyectos de minería de datos.

El énfasis especial en CRISP-DM obedece a que para tener un conocimiento claro de la minería de datos es más conveniente contar con una guía práctica que tener a mano conceptos teóricos de minería de datos que por sí solos no resultan útiles. Por lo tanto, se hace una explicación detallada del funcionamiento de CRISP-DM en el siguiente sub-capítulo.

5.2.6. CRISP-DM [2]

El proceso estándar transversal a la industria para minería de datos (Cross Industry Standard Process for Data Mining, CRISP-DM) es la metodología más usada en los grupos de interés de minería de datos. Comprende toda una base o marco de trabajo para realizar proyectos de minería de datos. Su éxito y alta aceptación se explican debido al enfoque que tuvieron los desarrolladores de CRISP-DM. No se construyó de manera teórica a partir de estudios académicos, sino desde la práctica, de las experiencias del mundo real de cómo se realizan proyectos de minería de datos.

CRISP-DM maneja una descomposición jerárquica para entender el proceso a través de distintos niveles de abstracción que se pueden clasificar así (Desde lo general hasta lo específico): Fase, tarea genérica, tarea especializada e instancia de proceso.

En CRISP-DM, un proyecto de minería de datos se divide en seis fases, compuestas a su vez de varias tareas genéricas. Este nivel se denomina genérico, porque debe ser lo suficientemente general para cubrir todas las situaciones de minería de datos posibles. Las tareas especializadas hacen referencia a cómo se deben realizar las tareas genéricas en situaciones específicas. A nivel de instancia de procesos se maneja un registro de las acciones, decisiones y resultados de un proyecto de minería de datos real.

Para entender mejor cómo se organiza un proyecto en CRISP-DM es indispensable describir las seis fases que lo componen y entender de entrada que no debe verse como una secuencia de pasos rígidos, sino que dependiendo del resultado de una fase o tarea específica se puede proceder hacia cierta fase o tarea. A continuación se hace una descripción del modelo de referencia CRISP-DM.

5.2.6.1. Entendimiento del negocio

Durante esta fase se debe tener claridad de los objetivos del proyecto y los requisitos desde una perspectiva de negocio. Esa síntesis debe transformarse en un problema de minería de datos y un plan preliminar de cómo lograr los objetivos.

Se compone de las siguientes tareas:

- **Determinar los objetivos del negocio.** Es importante para el analista de datos saber qué es lo que quiere el cliente, desde la perspectiva del negocio. Saltar este paso puede incurrir en gastar recursos produciendo las respuestas correctas para las preguntas equivocadas.

- **Evaluar la situación.** Se hace una inspección más detallada sobre recursos, restricciones, y otros factores a considerar al momento de determinar la meta del análisis de datos y el plan de proyecto. Deben estar presentes en una lista organizada.

- **Determinar las metas de minería de datos.** Se explican los objetivos del negocio en términos técnicos.

- **Producir el plan de proyecto.** Se realiza una descripción del plan esperado para cumplir ambos grupos de metas (De minería de datos y de negocio). Dentro de lo que debe incluir el plan están una selección inicial de herramientas y técnicas.

5.2.6.2. Entendimiento de los datos

Se empieza con una recolección inicial de los datos para identificar problemas de calidad en los mismos, detectar grupos de datos interesantes a la hora de establecer hipótesis, entre otras cosas.

Se compone de las siguientes tareas:

- **Recolectar los datos iniciales.** Se debe tener acceso a los datos y recursos listados en la fase anterior. Si se va a trabajar con una herramienta específica que requiere que los datos sean cargados, este paso es el momento adecuado para hacerlo. Se debe tener en cuenta que si existen varias fuentes, la integración es un paso importante a la hora de cargar datos.

- **Describir los datos.** Se examinan las propiedades de los datos “en bruto” y se realiza un reporte de resultados: Número de registros, cantidad de datos, campos por tabla, etc.

- **Explorar los datos.** Comprende una tarea más técnica, ya que se preocupa por problemas propios de minería de datos: Distribución de atributos claves; relaciones entre atributos; agregaciones simples (Simple Joins); análisis estadísticos. Puede ser útil a la hora de refinar la descripción de datos de la tarea anterior, o prepara el terreno para la transformación de datos en tareas posteriores.

- **Verificar la calidad de los datos.** Surgen preguntas del tipo: ¿Los datos que se tienen cubren los casos requeridos?, ¿Están correctos o contienen errores, y si hay errores, qué tan comunes son?, ¿Hay valores perdidos en los datos?

5.2.6.3. Preparación de los datos

En esta fase se construye el set de datos final (Aquellos datos que se usarán en la fase de modelado) a partir de datos en bruto. Probablemente haya que realizar las tareas de esta fase varias veces y sin un orden prescrito.

Se compone de las siguientes tareas:

- **Seleccionar los datos.** Se decide el grupo de datos a usar en el análisis. Entre los criterios a tener en cuenta hay restricciones de tamaño, o tipo de dato; calidad; e incluso relevancia a la hora de cumplir las metas de minería de datos.

- **Limpiar los datos.** De acuerdo a las técnicas de análisis escogidas es posible que se requieran datos más “limpios” (De mejor calidad). Se puede escoger de entrada grupos de datos que cumplan las condiciones para ser analizados con la técnica elegida.

- **Construir los datos.** Se incluyen tareas de generación de datos nuevos a partir de los existentes (Atributos derivados, nuevos registros o valores transformados por atributos previos).

- **Integrar los datos.** Creación de nuevos registros a partir de mezclar registros o valores anteriores.

- **Formatear los datos.** Modificaciones de tipo sintáctico que podrían ser útiles a la hora de incluir los datos en la herramienta de modelado.

5.2.6.4. Modelado

En esta fase se aplican técnicas de modelado, con su respectiva parametrización hacia valores óptimos. Debido a que algunas técnicas tienen requerimientos en cuanto al formato de los datos, en ocasiones hay que volver a realizar las tareas de la fase de preparación.

Se compone de las siguientes tareas:

- **Seleccionar la técnica de modelado.** Aunque en la fase de entendimiento del negocio posiblemente ya se haya elegido una herramienta de modelado, aquí se selecciona una técnica de modelado más concreta y específica (Redes neuronales, regresión, árboles de decisión, etc). Si se eligen varias técnicas, se realiza la tarea por separado.

- **Generar el diseño de prueba.** Se debe tener un procedimiento que permita probar la calidad del modelo escogido, y para esto se puede requerir tener grupos de datos separados (De prueba y de entrenamiento).

- **Construir el modelo.** Se ejecuta la herramienta de modelado con el grupo de datos seleccionado, lo que puede dar lugar a uno o varios modelos.

- **Evaluar el modelo.** Dependiendo del dominio de conocimiento se pueden tener criterios de éxito distintos, ya sea por el éxito de la aplicación del modelado o por los resultados de minería de datos desde la perspectiva del negocio. Si se tienen varias técnicas de modelado, es usual que se comparen los resultados.

5.2.6.5. Evaluación

A esta altura debe haber como mínimo un modelo construido, con una cierta calidad. En esta fase se evalúa exhaustivamente el modelo y se revisan los pasos ejecutados al construir el modelo, en aras de que cumpla los objetivos del negocio. Usualmente se revisan aquellos problemas de negocio que no se han tomado en cuenta lo suficiente.

Se compone de las siguientes tareas:

- **Evaluar resultados.** Aunque algunas tareas de fases previas contienen actividades de evaluación, se centran más en generalidades propias del modelo. Esta tarea busca ver si el modelo es deficiente y en qué situación del negocio. Además, se pueden revisar los resultados de minería de datos que no cumplen los objetivos del negocio originales, pero que pueden ser útiles en otros proyectos.

- **Proceso de revisión.** Se hace una revisión más profunda para detectar si se ha pasado algo por alto. Además, hay actividades propias de aseguramiento de calidad.

- **Determinar los siguientes pasos.** Según los resultados de las tareas anteriores se decide cómo proceder, ya sea dar paso al despliegue de resultados; realizar nuevas iteraciones; o dar paso a nuevos proyectos de minería de datos. Se pueden tener en cuenta temas de presupuesto y recursos remanentes.

5.2.6.6. Despliegue

Una vez que el modelo está creado es necesario organizar y presentar de manera adecuada el conocimiento adquirido al cliente. Esta fase puede variar de generar un simple reporte o ser tan compleja como implementar un proceso de minería de datos repetitivo. En ocasiones quien realiza el despliegue es el cliente mismo, que además necesitará entender qué acciones llevar a cabo para usar los modelos creados.

Se compone de las siguientes tareas:

- **Despliegue del plan.** De acuerdo a los resultados de evaluación se diseña una estrategia de despliegue. Si se reconoce un procedimiento genérico para crear modelos relevantes, se documenta para ocasiones posteriores.

- **Planificar el monitoreo y mantenimiento.** Es una tarea importante si los resultados del proyecto de minería de datos son de uso cotidiano para el negocio. El plan de monitoreo tiene en cuenta el tipo de despliegue específico.

- **Producir el reporte final.** El equipo del proyecto escribe un reporte final. Dependiendo del plan de despliegue, puede ser solamente un resumen del proyecto o una presentación final a fondo de los resultados.

- **Revisar el proyecto.** Se evalúa lo que estuvo bien, lo que estuvo mal, lo que se puede mejorar.

A pesar de la descripción que se hace de las tareas que componen las fases, no se toman en cuenta las salidas que se deben producir al final de cada una de ellas. Además de otros detalles que se pasan por alto debido al alcance de este documento. Por lo tanto, para un mayor entendimiento y desglose del modelo de referencia CRISP-

DM, se recomienda la lectura del documento **“CRISP-DM 1.0 Step-by-step data mining guide”**, compilado por la empresa **SPSS Inc.** (Uno de los fundadores del convenio que creó CRISP-DM), que también cuenta con una guía de usuario, la cual se diferencia en ciertos aspectos del modelo de referencia descrito anteriormente.

5.2.7. Algunos algoritmos de minería de datos

La mayoría de técnicas de minería de datos se basan en aprendizaje a partir de inducción, en el cual se crea el modelo de manera explícita o implícita generalizando a partir de un número adecuado de muestras de entrenamiento. La suposición que subyace al enfoque inductivo es que el modelo entrenado sea aplicable al futuro, sin que tenga que ver alguna muestra de ejemplo. Es decir, la inducción implica que las conclusiones no surjan a partir de deducción de las premisas.

Los métodos de minería de datos pueden clasificarse en dos grupos grandes de acuerdo a la taxonomía propuesta por Fayyad y otros. Métodos orientados a la verificación (En donde se verifica la hipótesis del usuario) y métodos orientados al descubrimiento (El sistema encuentra nuevas reglas y patrones de manera autónoma). Debido al alcance del documento, no se indagará en los métodos de verificación.

Por su parte, los métodos de descubrimiento se subdividen en dos grupos, los métodos de descripción y de predicción. Dentro de los métodos de descripción podemos ver el Agrupamiento (Clustering), Visualización, Sumarización, entre otros.

Los métodos de predicción comprenden un grupo más grande, diversificándose a través de dos ramas: Clasificación y Regresión**.** Dentro del primer subgrupo de clasificación encontramos algunos de los métodos más vistos en la literatura: Redes neuronales, Redes bayesianas, Árboles de decisión, Máquinas de vectores de soporte, y otros más.

Debido a la amplia variedad de métodos usados para minería de datos, se buscó sesgar la muestra. De acuerdo a una encuesta realizada en la comunidad virtual **KD Nuggets** [20], en la que se pregunta a los usuarios cuáles son los métodos/algoritmos que más usaron en análisis de datos en el último año (2011), los tres primeros lugares son para métodos de descubrimiento de distintas subclases:

- Árboles/Reglas de decisión.
- Regresión.
- Agrupamiento.

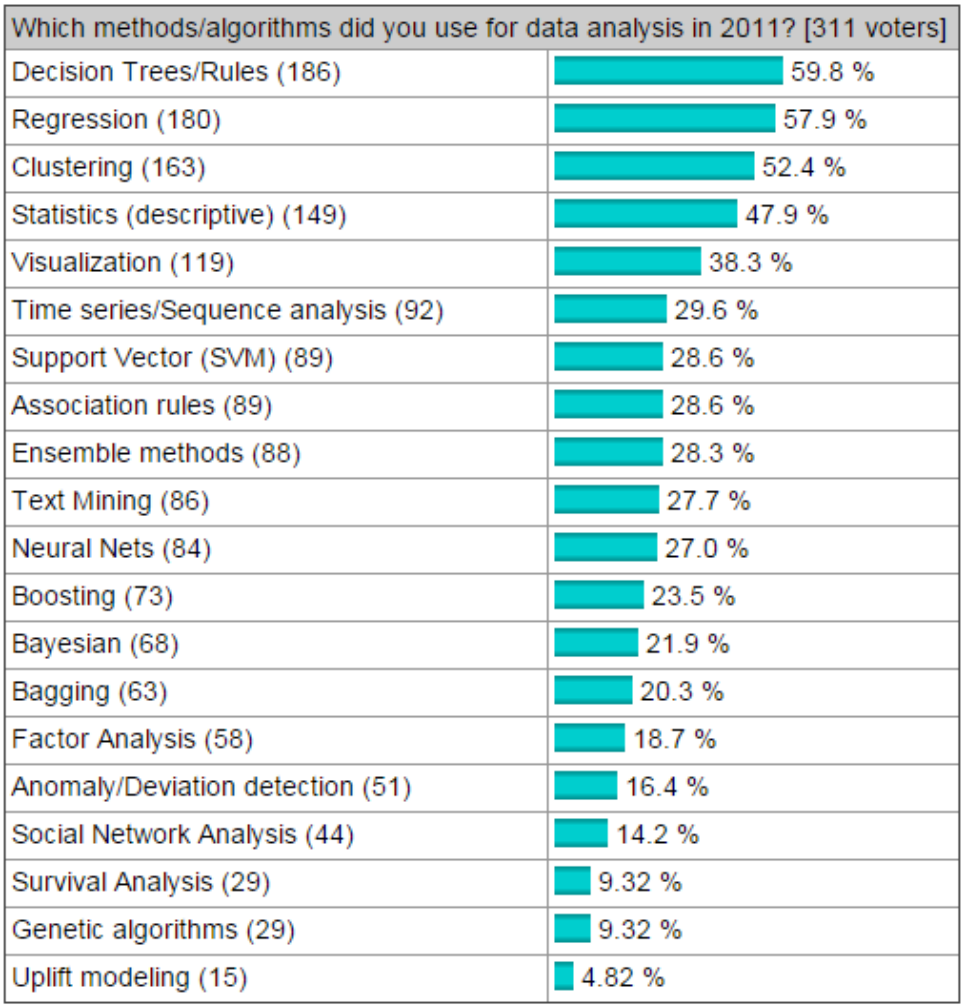

Ilustración 1: Algoritmos DM más usados

CAPITULO III

6. Estado del arte

6.1. Un sistema de soporte a la decisión basado en GIS para prevención de inundaciones en Quanzhou City. [28]

El control de inundaciones y reducción de desastres es una misión importante para garantizar el desarrollo económico nacional y social. Con la propuesta de “Hidrología y cuenca digital”, el sistema de información de prevención de inundaciones se vuelve una nueva tendencia en el campo de la ingeniería hidráulica. Tomando el sistema de información de prevención de inundaciones de Quanzhou como ejemplo, se discute el diseño e implementación del sistema de soporte basado en GIS para control de inundaciones en este artículo. Este sistema basado en el sistema de información de hidrología crea relaciones de precipitaciones en tiempo real y búsqueda de flujos, administración de control de inundaciones, predicción y simulación de inundaciones, diseminación de información de inundaciones, evaluación de daños relacionados, etc. Este artículo ilustra la arquitectura del sistema, el marco de trabajo funcional y el diseño del subsistema de soporte a la decisión. Como una medida de no-ingeniería, este sistema de soporte a la decisión juega un papel vital en el control de inundaciones y reducción de desastres, que es una parte irremplazable de todo el sistema de control de inundaciones.

Basado en un sistema de información geográfica combinado con inteligencia artificial, recolección de información y una gran cantidad de técnicas computacionales este paper nos muestra una de las formas de usar un sistema de soporte a la decisión por fuera el ámbito comercial. Creando un sistema de pronóstico de inundaciones el cual puede predecir la escala de afectación de la misma, brindando la habilidad de una toma de decisiones más rápida y efectiva ante una situación de este estilo.

6.2. Inteligencia industrial - Un enfoque basado en inteligencia de negocios para mejorar la ingeniería de fabricación en compañías industriales. [29]

La flexibilidad, la eficiencia de los recursos, y el tiempo de salida al mercado son factores claves de éxito para las empresas industriales. Las configuraciones esenciales se hacen durante las fases tempranas del desarrollo del producto así como de la fabricación. En las últimas fases del ciclo de vida del producto, las respuestas del mercado (Por ejemplo, quejas o cantidad de casos de daño) muestran la etapa de madurez de estos productos. Los métodos de calidad como TQA o EFQM persiguen el objetivo de aprender permanentemente de esta información. Por lo tanto es necesario tener un suministro de información adecuado. Este artículo se enfoca en este problema en el contexto de la gestión de la etapa de madurez en ingeniería de fabricación. Por consiguiente la investigación identifica primero un enorme vacío entre la fuente de información discutida teóricamente, basada en abarcar bases de datos, y los escenarios IT heterogéneos reales, que han crecido en la historia. En base a hallazgos empíricos, los negocios industriales carecen de conceptos que los ponen en una posición de adecuados suministros de información. Por lo tanto, un concepto de Inteligencia de Negocios genérica, desarrollado a través de actividades de investigación, parece ser un enfoque prometedor. Así es posible combinar información de los rasgos del producto e información de la fabricación con las dimensiones tradicionales del análisis gerencial, en aras de identificar el impacto de las decisiones de ingeniería en el ciclo de vida del producto.

Este es un ejemplo claro de la importancia que tiene la implementación de una herramienta de inteligencia de negocios para la toma de decisiones en una empresa, ya que al no contar con una herramienta BI, la empresa tiene acceso a una limitada (y en algunos casos irreal) cantidad de información, Se muestra como la implementación de una herramienta de BI convencional, hace más fácil la labor gerencial en la toma de decisiones, con lo cual se podrá alcanzar el objetivo puntual del texto que es la gestión de la etapa de madurez de los productos.

6.3. Propuesta de modelo para procesos ETL de bodega de datos. [30]

Las herramientas de Extracción - Transformación - Carga (Extraction - Transformation - Loading, ETL) son piezas de software responsables por la extracción de datos de diferentes fuentes, su limpieza, personalización, reformateo, integración e inserción hacia una bodega de datos. Construir el proceso ETL es potencialmente una de las tareas más grandes a la hora de construir una bodega de datos; es complejo, consume buen tiempo, y consume la mayoría de los esfuerzos, costos y recursos de implementación de una bodega de datos. Construir una bodega de datos requiere enfocarse muy de cerca en entender tres áreas principales: El área de la fuente, el área de destino, y el área de mapeo (Procesos ETL). El área de fuente tiene modelos estándar como el diagrama entidad-relación, y el área de destino tiene modelos estándar como el esquema de estrella, pero el área de mapeo no tiene un modelo estándar hasta ahora. A pesar de la importancia de los procesos ETL, se le ha dedicado poco a la investigación de esta área debido a su complejidad. Hay una clara carencia en el modelo estándar que se puede usar para representar los escenarios ETL. En este artículo se tratará de navegar a través de los esfuerzos realizados para conceptualizar los procesos ETL. La investigación en el campo de modelado de procesos ETL se puede categorizar en tres enfoques principales: Modelado basado en expresiones y pautas de mapeo, modelado basado en construcciones conceptuales, y modelado basado en ambiente UML. Estos proyectos tratan de representar las principales actividades de mapeo a nivel conceptual. Debido a la variedad de diferencias entre las soluciones propuestas para el diseño conceptual de procesos ETL y debido a sus limitaciones, este artículo también propondrá un modelo para el diseño conceptual de procesos ETL. El modelo propuesto se basa en la mejora de modelos previos para soportar algunos rasgos de mapeo que no se tienen en cuenta.

Los procesos ETL son el corazón de la construcción de las bodegas de datos. Se relacionan con el aprovisionamiento de los datos hacia las bodegas de datos. Todo el proceso comprende desde la extracción de los datos (Ya sea desde fuentes

transaccionales, como una base de datos, o desde fuentes no transaccionales, como un archivo de texto plano); pasa por la transformación de los datos, cuyo resultado se guarda en un área de almacenamiento y se transforma nuevamente de manera iterativa para ir adquiriendo datos más limpios y precisos; y finaliza con el cargue de datos en la bodega de datos como tal.

La propuesta que se realiza en este artículo es denominada Diagrama de Mapeo de Entidad (Entity Mapping Diagram, EMD) y su funcionamiento se puede resumir en el cumplimiento de los siguientes seis requisitos:

- Soporta la integración de múltiples fuentes.
- Es robusto al observar fuentes de datos cambiantes.
- Soporta transformaciones flexibles.
- Se puede desplegar fácilmente en un ambiente de implementación adecuado.
- Es lo suficientemente completo como para soportar varias operaciones de ETL.
- Es simple de crear y mantener.

6.4. Minería de datos - Pasado, presente y futuro - Un estudio típico en haces de datos [31]

La minería de haces de datos es una de las áreas que está ganando significancia práctica y progresa a pasos agigantados con nuevos métodos, metodologías y hallazgos en varias aplicaciones relacionadas con la medicina, las ciencias computacionales, bioinformática, y precisión del mercado bursátil, predicción del clima, procesamiento de vídeo, texto y audio, entre tantos otros. Los datos pasan a ser una preocupación clave en la minería de datos. Con los enormes datos en línea generados de varios sensores, salas de Chat, Twitter, Facebook, bancos en línea o transacciones de cajeros, el concepto de datos dinámicamente cambiantes es un reto importante, conocido como haces de datos. En este artículo, se provee el algoritmo para hallar

patrones frecuentes de haces de datos con un caso de estudio y se identifican los problemas de investigación al manejar haces de datos.

Los haces de datos (DS, Data Streams) pasan a convertirse en un término clave para entender el futuro de la minería de datos teniendo en cuenta la velocidad con la que se generan datos en la actualidad. Cabe aclarar que cuando se aplica minería de datos sobre un grupo de datos, éste puede tener un comportamiento estático (Está disponible para el análisis antes de su procesamiento y generalmente no varía con el tiempo) o dinámico (Altamente voluminosos y generalmente cambiantes con el tiempo, no permitiéndoles estar disponibles para procesamiento o análisis). Dentro del último grupo se pueden clasificar los haces de datos.

Algunos factores hacen que el procesamiento de haces de datos en la actualidad siga siendo un reto y objeto de estudio para varios investigadores. Haciendo una comparación entre el procesamiento de bases de datos (DB, Data Bases) y el procesamiento de haces de datos podemos tener en cuenta los siguientes rasgos: En DB el acceso a los datos se puede realizar de manera secuencial o no, en DS es siempre secuencial; los resultados computacionales en DB son precisos, mientras que en DS suelen hallarse aproximaciones; la velocidad de llegada de los datos se puede ignorar en DB, pero en DS la tasa de llegada de datos es mayor que la tasa de procesamiento. Estos son sólo una muestra de las características que sustentan el panorama de los haces de datos y su campo de investigación.

6.5. Prediciendo fallas de negocio usando Árboles de Regresión y Clasificación: Una comparación empírica con métodos estadísticos populares clásicos y métodos de minería de clasificación de alto nivel [32]

Predecir fallas de negocio es una tarea muy crítica para oficiales de gobierno, accionistas, gerentes, empleados, inversionistas e investigadores, especialmente en un ambiente de competencia económica como el de hoy. Algunos métodos de minería de datos dentro de los diez principales se han vuelto alternativas muy populares en la

Predicción de Fallas de Negocio (Business Failure Prediction, BFP), como por ejemplo las Máquinas de Soporte a Vectores (Support Vector Machine, SVM) y El Vecino-K-Más-Cercano. En comparación con otros métodos de minería de clasificación, las ventajas de los Árboles de Clasificación y Regresión (CART) incluyen: Simplicidad de resultados, fácil implementación, estimación no-lineal, es no-paramétrico, preciso y estable. Sin embargo, no hay muchos investigadores en el área de BFP que atestigüen la aplicabilidad de CART, otro método entre los diez mejores en la minería de datos. El objetivo de este artículo es explorar el rendimiento de BFP usando CART. Para demostrar la efectividad de BFP usando CART, se desarrollaron tareas BFP en el set de datos recolectados de compañías listadas en el Shanghai Stock Exchange y el Shenzhen Stock Exchange. Se empleó el método treinta veces como evaluación, además de los métodos de SVM y vecino-k-más-cercano, y los dos métodos de referencia del área estadística, es decir, Análisis Múltiple Discriminante (Multiple Discriminant Analysis, MDA) y regresión logísticas, que se emplearon como métodos comparativos. Para los métodos comparativos, se empleó el método paso a paso de MDA y así seleccionar un subgrupo de rasgos óptimo. Los resultados empíricos indicaron que el algoritmo óptimo de CART supera todos los métodos comparativos en términos de rendimiento predictivo y pruebas de significancia en BFP de corto-alcance de las compañías chinas listadas.

Un estudio comparativo hecho de manera empírica pero con una metodología clara y sistemática permite visualizar el papel de algunos de los algoritmos de minería de datos de más uso, haciendo un énfasis en un área de negocio crítico para ciertos sectores. Tomar esto en consideración permitirá seleccionar la técnica adecuada a la hora de abordar este tipo de problemas, mostrando además las virtudes de los demás algoritmos, que aunque parezcan “insuficientes” o con un rendimiento menor en los problemas BFP, pueden llegar a ser útiles en estudios más académicos o investigativos.

6.6. El árbol de decisión CART para haces de minería de datos [33]

Una de las herramientas más populares para minería de haces de datos son los árboles de decisión. En este artículo se propone un nuevo algoritmo, que está basado en el algoritmo comúnmente conocido como CART. La tarea más importante al construir árboles de decisiones para haces de datos es determinar el mejor atributo para hacer un corte en el nodo considerado. Para resolver este problema se aplica aproximación Gaussiana. El algoritmo presentado permite obtener alta precisión de clasificación con un tiempo de procesamiento corto. El principal resultado de este artículo es el teorema que muestra que el mejor atributo calculado en el nodo considerado según la muestra de datos es el mismo, con una alta probabilidad, como el atributo derivado de todo el haz de datos.

Se presenta una propuesta para procesamiento de haces de datos dinámicos usando como base el concepto CART. El procesamiento de haces de datos se presenta como un campo de investigación novedoso al intentar analizar datos que se generan a una velocidad mayor, teniendo serias restricciones en cuanto a la capacidad de memoria, la cantidad de escaneos que se pueden realizar sobre el grupo, entre otras. CART para haces de datos (dsCART) surge como una modificación del algoritmo de Árboles de Decisión Muy Rápidos, introduciendo un mecanismo de tie breaking, que esencialmente lo que hace es establecer un criterio de corte en cierto nodo a partir de un número de elementos fijos.

6.7. Clasificación jerárquica para Bodegas de Datos: Un estudio [34]

En los sistemas de bodegas de datos, la jerarquía juega un rol clave al procesar y monitorear información. Estas jerarquías analizan dinámicamente altos volúmenes de datos históricos en bodegas de datos de varios niveles de granularidad usando operaciones OLAP como roll-up y drill-down. A través de estas operaciones se pueden resumir y detallar los datos que ayudan al análisis y el proceso de toma de decisiones.

Algunos autores han definido jerarquías que derivan de aplicaciones del mundo real para representar un amplio rango de escenarios de negocios. Pero hay una necesidad de categorizar propiamente jerarquías de dimensión así como modelarlas adecuadamente durante la evolución. En este artículo se ha provisto una comparación comprensiva de diferentes categorías de jerarquías propuestas por varios investigadores basados en ciertos parámetros.

Es claro que la alta comprensibilidad de una bodega de datos es gracias a su división implícita en hechos y dimensiones, términos que pueden ser entendidos incluso a altos niveles por usuarios gerenciales. Pero existen tantos tipos de negocios, y tantos tipos de gerentes ligados a éstos últimos, que es necesario establecer una jerarquía genérica y transversal a la mayoría. Las jerarquías definidas en algunos atributos de dimensión son principalmente esenciales si son capaces de enfrentar los problemas en el análisis de datos. Las jerarquías de dimensión permiten observar los datos a distintos niveles de granularidad, iniciando con una vista general hasta una detallada a través de la operación de Drill-Down. Si se quiere el camino inverso, es útil la operación de Roll-Up.

Para dar una introducción básica y técnica, una jerarquía se define como un grupo de relaciones binarias existentes entre dos niveles de dimensión, conocidos simplemente como Niveles. Teniendo dos niveles consecutivos, el nivel más alto se conoce como Padre y el nivel más bajo se conoce como Hijo. La estructura que representa una jerarquía es lo que importa para el análisis. Soportar varios tipos de jerarquías y tener definiciones flexibles de las mismas permite tener un rango más amplio de escenarios a modelar.

6.8. Integración de BI y ERP dentro de las organizaciones [35]

Las organizaciones han invertido una cantidad considerable de recursos en la implementación de sistemas de Planeación de Recursos Empresariales (ERP) e Inteligencia de Negocios (BI). En el ambiente competitivo de hoy dentro del contexto de

BI y ERP complejos, estos sistemas se han vuelto herramientas estratégicas claves, las cuales impactan directamente en el éxito de cualquier implementación de proyecto. Pero se le ha dado poca atención a la integración de Inteligencia de Negocios y Planeación de Recursos Empresariales (BIERP). Se han llevado toda una serie de estudios en los países desarrollados mientras que aquellos en desarrollo recibieron mucha menos atención. A pesar del esfuerzo que se ha realizado para explicar la integración de estos sistemas, la literatura todavía se califica como fragmentada y diversa. Este artículo intenta revisar y evaluar artículos publicados entre 2000 y 2012 relacionados con la integración BI-ERP.

La implementación de herramientas BI y de un sistema ERP en las compañías ya es un estándar a nivel mundial. Sin embargo, aún son pocas las empresas que se preocupan por integrar estas dos herramientas. Un sistema ERP permite tener todos los datos de la empresa en un solo lugar, disponibles para todas las áreas de la misma. Sin embargo, la idea principal de un ERP no es el de brindar reportes o indicadores en línea. Es aquí donde entra a jugar parte importante una solución de BI.
Una herramienta de BI me permite manejar reportes, indicadores, tableros de control y todo tipo de información que necesite cualquier persona sobre cualquier tema en la compañía. Sin embargo, la herramienta por sí misma no obtiene esta información, necesita un suministro de datos para poder llevar a cabo todo el proceso de transformación de datos en información y en previo conocimiento para la empresa.
Este artículo nos presenta una cantidad de información recopilada, explicando el por qué las empresas deben buscar la integración de un sistema ERP y una herramienta de BI (BIERP). Planteando el ERP como suministro de datos e información y la solución de BI como herramienta administrativa que brinda el conocimiento necesario para ejecutar la toma de decisiones en una empresa.

6.9. Inteligencia de Negocios Social: Una nueva perspectiva para los encargados de tomar decisiones [36]

Este artículo presenta un enfoque de Inteligencia de Negocios Social, una gestión teórica que se soporta sobre un ejemplo práctico. El marco de trabajo general para definir la perspectiva social está sujeto al presente debate, mientras los datos sociales se modelan dentro de un esquema de bodega de datos. Ciertamente, los datos sociales no son la única fuente de datos de la cadena de valor de BI, pero define una nueva perspectiva para los encargados de tomar decisiones.

Hoy en día, cualquier tipo de compra, cotización, búsqueda en internet y un sinfín de movimientos más; entrega información valiosa de gustos y tendencias a las compañías que ofertan bienes y servicios. Sin sumar a esta información (la cual se brinda inconscientemente), la información que es recolectada en encuestas, buzones de sugerencia y demás mecanismos de captura de información; es bastante como para que una empresa pueda ejercer planes de venta en los compradores comunes de su negocio.

Aprovechar la predisposición de la sociedad a brindar información de sus preferencias y gustos, es una de las estrategias para la recolección de información que llevan a cabo las grandes empresas. Recopilando esta información y almacenándola en bodegas de datos, puede ser usada después como fuente de conocimiento para la toma de decisiones.

6.10. Agentes computacionales para DSS [37]

En los sistemas de soporte a la decisión, es esencial tener una solución candidata rápida, incluso si hay que recurrir a una aproximación. Esta restricción introduce un requisito de escalabilidad que tiene en cuenta el tipo de heurística que se puede usar en tales sistemas. Cuando el tiempo de ejecución es limitado, estos algoritmos necesitan dar mejores resultados e incrementar los recursos computacionales en vez de tiempo adicional. En este artículo se muestra cómo los sistemas multi-agentes pueden llenar estos requisitos. Se invoca un ejemplo del concepto de Sistemas Multi-Agentes Evolucionarios, que combinan paradigmas evolucionarios y de agentes

computacionales. Se describen varias implementaciones posibles y se presentan resultados experimentales que demuestran cómo los recursos adicionales mejoran la eficacia de tales sistemas.

Los sistemas multi-agentes evolutivos (Evolutionary Multi-Agents Systems, EMAS) son una metaheurística que basa su funcionamiento en dos corrientes actuales en la computación, la computación basada en agentes y la computación evolutiva. La implementación de este tipo de sistemas puede ser más eficiente de manera secuencial y optimizada que al asignar un agente a un hilo sin orden alguno. La escalabilidad de un EMAS está ligada directamente a los recursos computacionales disponibles.

Además de tales corrientes, un concepto clave en EMAS es el de las interacciones inteligentes, como la coordinación, la cooperación y la negociación. También se debe tomar en cuenta que los diversos agentes que colaboran en el sistema lo hacen de manera autónoma, cumpliendo las tareas asignadas de acuerdo a una estrategia propia y a su visión del ambiente. Sin embargo, hay que prestar atención suficiente a la coordinación de las actividades de los distintos agentes.

6.11. Un sistema de soporte a la decisión para el diseño y gestión de sistemas de almacenamiento [38]

El problema del manejo material involucra el diseño y control operativo de sistemas de almacenamiento (Por ejemplo, centros de distribución), que permiten emparejar vendedores y solicitudes, arreglar por temporadas, consolidar productos y agrupar actividades de distribución. Los sistemas de almacenamiento jugan un papel crucial en proveer eficiencia y satisfacción al cliente. El diseño de almacén implica un amplio grupo de decisiones, que involucran restricciones de disposición y problemas operativos que afectan seriamente los rendimientos y los costos logísticos totales.

Este estudio presenta un sistema de soporte a la decisión original para el diseño, gestión y control de sistemas de almacenamiento. Específicamente, el DSS propuesto implementa una metodología Top-Down que considera el diseño estratégico de almacenes y la gerencia de operaciones. El DSS puede simular los rendimientos de logística y manejo de materiales de un sistema de almacenamiento. los métodos y algoritmos heurísticos dirigen varios problemas críticos de los almacenes, como el orden en el proceso de elección, el cual es responsable del 55% de los costos totales en un centro de distribución. Los beneficios debido a la adopción del DSS propuesto se resumen en una tabla de Indicadores de Rendimiento Claves (*Key Performance Indicators*) de eficiencia en espacio y tiempo que permiten a los proveedores de logística, practicantes y gerentes así como académicos y educadores enfrentar instancia de almacenamiento en el mundo real y encontrar lineamientos útiles para el manejo de materiales.

Aprovechando las ventajas del músculo de procesamiento que provee el marco de trabajo de un DSS se pueden realizar implementaciones para la optimización de procesos y reducción de costos en situaciones del mundo real. La sistematización del manejo de un almacén o bodega representa un hecho importante para administradores y empleados en un centro de distribución, dado por el mejoramiento de la logística en el manejo de los productos. Tiempo y espacio se pueden optimizar como dos medidas trascendentales, pero el proceso de elección de productos se considera un factor más importante a tener en cuenta. Configurar el escenario de manera adecuada para que el DSS resuelva este tipo de problemas es el verdadero reto, lo que incluye modelar la lógica del negocio para que ésta se adhiera a detalles técnicos.

CAPITULO IV

7. Arquitectura de una solución de Business Intelligence

Una solución de Business Intelligence permite tener una mejor visualización del entorno de todo el negocio, generando un valor agregado no sólo para las estrategias comerciales, sino para la compañía en general. Procesos de producción, operaciones financieras, administración de un CRM y todos los demás procesos de la compañía adquieren una ventaja con respecto a los procesos de otras empresas que no cuentan con un servicio de este tipo.

Cuando se habla de la implementación de una herramienta o solución de BI, se abarca un área de trabajo muy extensa, la cual puede dividirse en 5 ejes principales: Captación de la información, Manejo de la información, Visualización y distribución, Análisis de la información y Gestión de las decisiones adoptadas.

Cada uno de estos ejes comprende una etapa de la transformación de los datos y la información en conocimiento, lo que es la esencia de las herramientas de BI, y para cada una de esas etapas es necesario tener una herramienta tecnológica que nos soporte el almacenamiento, transformación, visualización, y demás acciones que se deban realizar sobre la materia prima de BI, los datos.

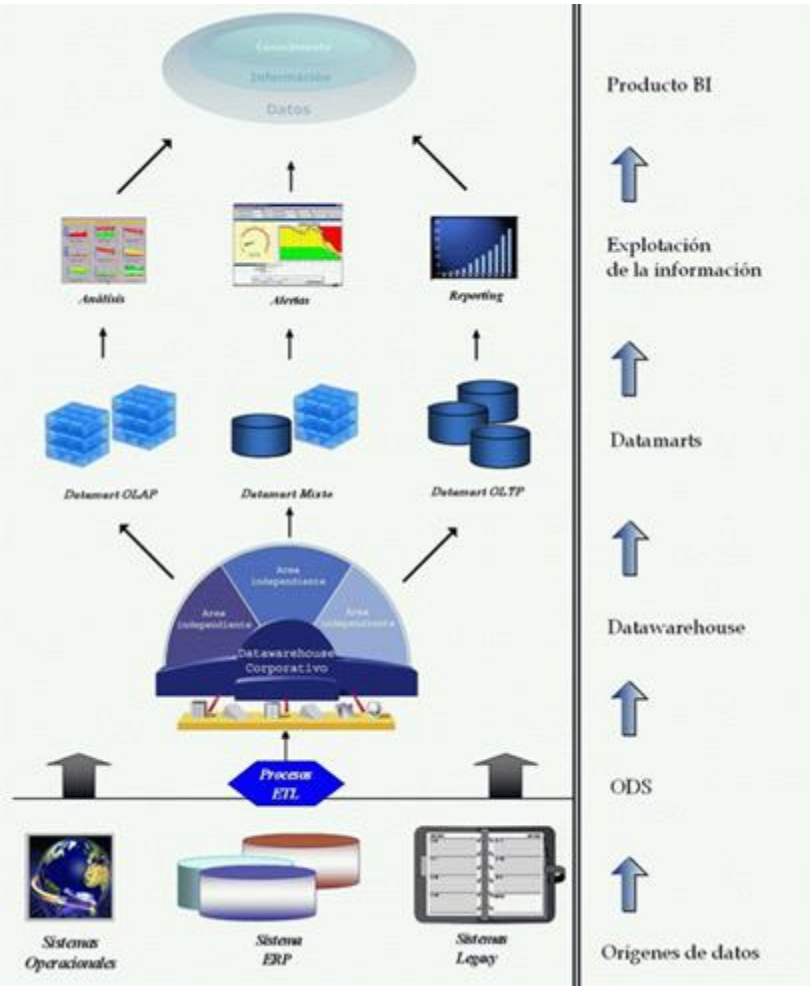

Ilustración 2: Arquitectura BI

7.1 Captación de la información

Este eje comprende todo lo que tiene que ver con la integración de los datos. En la gran mayoría de los casos, cuando se implementa una solución BI nos encontraremos con información dispersa, proveniente de distintos orígenes de datos: sistemas ERP, sistemas operacionales, bases de datos, hojas de cálculo, archivos planos, entre otras. Al proveer de diferentes orígenes, los datos pueden tener problemas de integridad, calidad, estructura u orden. En esta etapa de la implementación, BI se ve apoyado en las herramientas ETL, abreviado así por sus siglas en ingles Extract, Transform and Load.

Extraer, transformar y cargar. Nombre que describe perfectamente la funcionalidad de este tipo de herramienta, la cual está encargada de la primera interacción de los datos de una empresa con el proceso de implementación de una solución BI. Una herramienta ETL sigue un conjunto de pasos lógicos, que inicia por la extracción de los datos de sus diferentes orígenes. Los datos extraídos son analizados para garantizar que cumplan la estructura necesaria para el siguiente paso, de no cumplir con el estándar, los datos son rechazados. La extracción de estos datos generalmente se realiza desde bases de datos o archivos planos en producción, por lo cual el proceso se debe realizar en horarios programados que no afecten el correcto funcionamiento de la compañía. Una vez se extraen los datos, estos son consignados en una especie de plantilla o formato, el cual se entregara al siguiente paso de los ETL, la transformación.

En la etapa de transformación la mayoría de los datos extraídos sufren un pequeño ajuste, con lo cual se garantizará que todos los datos que se obtienen tengan la misma estructura. Estos pequeños cambios pueden ser: cambios de tipo de datos, dividir datos de una columna en otras, crear campos nuevos a partir de cálculos, eliminar columnas con información irrelevante, codificar campos y demás. Una vez se hayan realizado los cambios necesarios, la herramienta ETL debe validar que cada uno de los datos cumplan con la estructura necesaria, de no ser así estos se descartan completa o parcialmente, y si se encuentran corregidos pasaran a la última etapa del proceso.

La última etapa de este proceso hace referencia a la carga de los datos en el nuevo sistema, el cual puede ser un Data Warehouse o diferentes Data Marts, dependiendo de la necesidad de cada empresa. Al momento de cargar la información esta puede quedar almacenada a modo de resumen, guardando resultados de cálculos como una única transacción; o también se pueden tener varios niveles de granularidad, que me permite manejar más dimensiones de información, para tener información más detallada en sus diferentes jerarquías.

7.2 Manejo de la información

Una vez se cuenta con información unificada, integra y consolidada se da paso a la siguiente etapa o eje de la implementación de BI, en la cual se lleva a cabo la administración de los datos y la información. En este eje las herramientas tecnológicas que apoyan el proceso son las de almacenamiento de datos, Data Warehouse o Data Marts. Estos almacenes de datos cuentan con la estructura indicada para un óptimo análisis de los datos de la empresa, ya sean bases de datos de tipo OLAP u OLTP.

Las bases de datos de tipo OLAP (abreviado así por sus siglas en ingles “On-Line Analytical Processing”), como su nombre lo indica, son bases de datos orientadas al procesamiento analítico. Este tipo de base de datos permite tomar grandes volúmenes de datos y presentarlos como información más compacta y útil: resúmenes de tendencias en las ventas, mejores compradores, distribución de las ventas por horarios, y demás reportes que sean utilizados por las diferentes áreas de la empresa.

El nombre “On-Line Transactional Processing”, más conocido por sus siglas OLTP, hace referencia a las bases de datos que están orientadas al procesamiento de transacciones. Cuando se habla de transacciones en una base de datos se hace referencia a procesos e inserción, modificación o eliminación de datos. Este tipo de DB cuenta con un acceso optimizado a sus datos, debido a las frecuentes consultas que se hacen sobre ellos.

En este eje vale la pena hacer referencia a otro almacén de datos que sirve de enlace entre el eje anterior y éste. Un almacén de datos ODS (Operational Data Store), el cual actúa como almacén de paso entre el origen de los datos y el almacén de destino final. Mientras los datos se encuentran en un almacén ODS, pueden ser modificados, eliminados o verificados. Todo esto para asegurarse de que los datos que pasen al almacén de destino tengan la mayor integridad posible. Este tipo de almacén está diseñado para consultas que demanden un bajo consumo de procesamiento.

7.3. Visualización y distribución

Si recordamos, el objetivo final de la implementación de una solución de BI es la de convertir los datos y la información en conocimiento, así que todos los datos que recopilamos y agrupamos en las dos etapas anteriores deben ser mostrados a los jefes de procesos y directivos de la compañía, los encargados de la toma de decisiones. La forma de presentar los datos a estos ejecutivos debe ser de resumida y lo más gráfica posible, para este fin contamos con algunas herramientas tecnológicas en la visualización de la información, las cuales, en esencia, nos permiten mostrar información de forma ordenada, clara y simple.

El "reporting" es una herramienta que permite mostrar la información de un proceso determinado de la compañía de forma clara y concreta. Por medio de gráficos de barras, gráficos circulares, distintos tipos de tablas, entre otros; se presenta la información al nivel gerencial de la compañía.

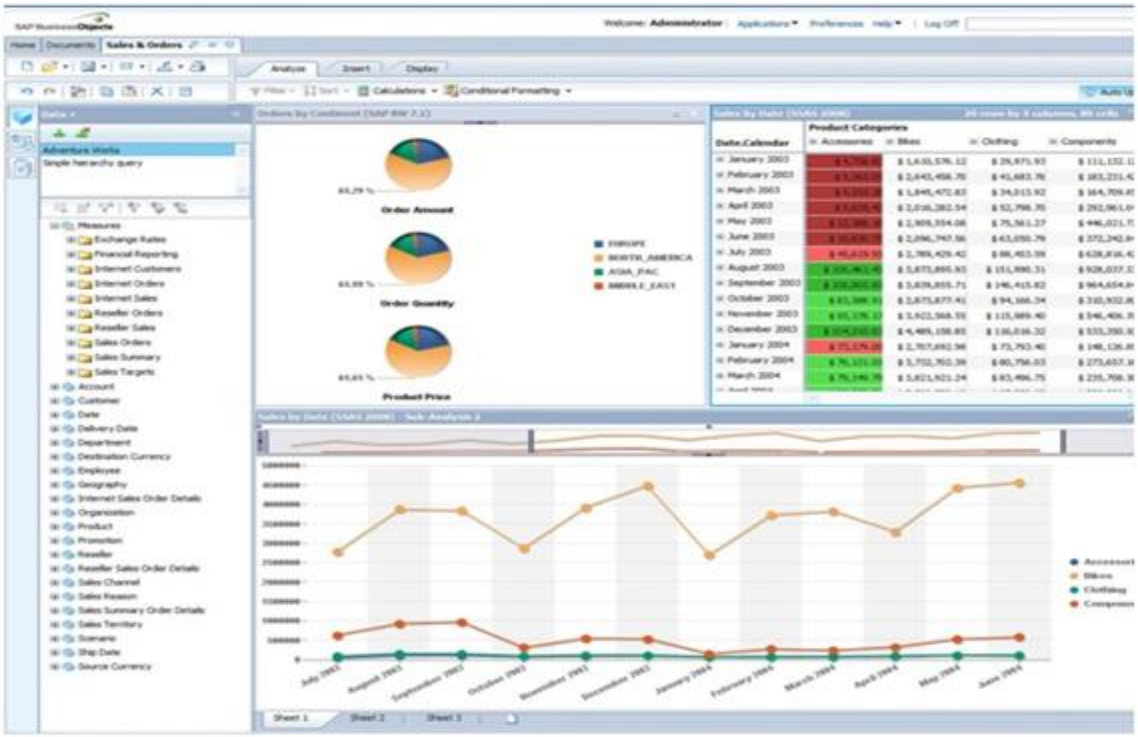

Ilustración 3: Reporting

Esta herramienta es la más común en este eje de la implementación. Fácil de crear, configurar y mantener, poca interactividad del usuario final, con campos parametrizables y predefinidos; son algunas de las ventajas que hacen que esta solución sea una de las preferidas por los miembros del área de tecnología en la empresa. Brindan toda la información necesaria, sin necesidad de estar realizando constantes cambios en la extracción de la información.

Otro tipo de visualización y distribución de los datos y la información son los “Dashboards”, otra de las herramientas más populares, siendo ésta un poco más compleja e interactiva que la anterior. Los “Dashboards” son interfaces de uma o más páginas que permiten acceder a la información por medio de alertas, gráficas, reportes, indicadores y otras técnicas de análisis. Debido al contacto más cercano que tiene el usuario final con los “Dashboards”, estos deben ser más amigables, personalizables e interactivos.

La información que alimenta los Dashboards está en los sistemas ERP, en la información capturada por los CRM o directamente de un Data Warehouse. En base a esta información se construyen indicadores, los cuales van a ser medidos a través de métricas que se definen en el mismo sistema. Un Dashboard me permite administrar alertas, las cuales se activan al transcurrir distintos eventos programados previamente en la herramienta. Por ejemplo, cuando hay una tendencia negativa en una de las métricas creadas anteriormente.

Al visualizar la información de la empresa desde un Dashboard se facilita la toma de decisiones en la compañía, ya que en él, se pueden observar claramente, cómo se comportan los indicadores clave de desempeño (KPI). Compartir, agrupar y centralizar los datos de la información en un solo lugar, estas son las ventajas de tener un Dashboard como herramienta de visualización de la información.

7.4. Análisis de la información

Hasta hace algunos años el proceso de Business Intelligence se consideraba sólo hasta este punto. La visualización de los datos recopilados como información valiosa para el comité gerencial de una compañía. De aquí en adelante todo lo debían hacer los integrantes de este comité, la toma de decisiones estaba basada en su experiencia como líderes o administradores, y tiene sentido. Las decisiones son actos puntuales que realizan los seres humanos basados en experiencia, gustos o suposiciones. Así que no se podía imaginar un avance tecnológico más allá de lo que se planteaba en este momento.

Sin embargo, la toma de decisiones en un contexto empresarial no es un hecho puntual, es el resultado de un proceso, y los procesos sí pueden ser gestionados por una herramienta tecnológica. Hoy en día estas tecnologías apuestan al almacenamiento inteligente, la generación de informes y los modelos de análisis predictivo. Siendo los últimos a los que más apunta el mercado en este momento.

Las compañías que actualmente poseen una herramienta de Business Intelligence poderosa, usan su capacidad de predicción para realizar la ubicación de nuevos puntos de venta, definir la cantidad de producción, definir las características de los productos, determinar descuentos acertados para los productos de baja rotación, entre muchas más decisiones. Las decisiones que sugiere un modelo de análisis predictivo lo realiza basado en todas las transacciones que tiene en la base de datos, patrones de comportamiento y manejo de la incertidumbre.

7.5. Gestión de las decisiones adoptadas

Una vez se toma una decisión, el siguiente paso es ponerla en marcha y realizar el correcto seguimiento a ésta. Es con este eje con el que se cierra el proceso de

implementación de una solución de Business Intelligence, el cual tiene como objetivo gestionar la totalidad del proceso de toma de decisiones en una compañía.
Para llevar a cabo la última etapa de este proceso, los líderes de procesos se pueden apoyar en una herramienta de gestión de proyectos o gestión de responsabilidades. Las cuales brindan el seguimiento necesario para cualquier tipo de proyecto que se esté planteando la empresa. Como lo son las decisiones tomadas en base a la el conocimiento brindado por una solución de Business Intelligence.
Son muy pocas las soluciones de BI que tienen implementada una herramienta de gestión de proyectos, pero este es el paso a dar de ahora en adelante. Con la implementación de una herramienta de este tipo en las soluciones BI se cerrará completamente el ciclo de transformación de los datos a información, de información a conocimiento y de este conocimiento adquirido en inteligencia.

8. Tipos de problemas de minería de datos y cómo abordarlos: Reglas/Árboles de decisión; Regresión; Agrupamiento

Un proyecto de minería de datos usualmente está compuesto por distintos tipos de situaciones problemáticas, que en conjunto constituyen el objetivo a atacar. Por lo tanto, conocer cuáles son en general estos problemas y cuáles son las alternativas a usar es parte importante a la hora de hablar sobre minería de datos. De acuerdo a la guía realizada por el Grupo de Interés Especial (SIG, por sus siglas en inglés) de CRISP-DM, las situaciones más comunes son:

8.1. Tipos de problemas de minería de datos [2]

8.1.1. Descripción de datos y resumen

El objetivo a alcanzar en este tipo de problema es el detalle conciso de las características de los datos, para proveer al usuario de una visión global de las estructuras de datos presentes. A menudo, la descripción de datos podría ser uno de

los objetivos generales de un proyecto de minería de datos, pero generalmente hace parte del proceso como una sub-meta en las etapas más tempranas. Esto se debe al desconocimiento usual de los datos y su naturaleza por parte de los usuarios. Una exploración inicial de los datos puede dar lugar a hipótesis potenciales en información escondida.

La descripción de datos y resumen típicamente se combinan con otros tipos de problemas. Por ejemplo, una descripción detallada de los datos puede llevar a la postulación de segmentos interesantes en los datos. Una vez identificados y definidos, es útil hacer una nueva descripción y resumen de estos segmentos. Eso sí, es recomendable que se aborde la descripción y resumen de los datos antes de dirigirse a otro tipo de problema. Incluso, se podría tomar este tipo de problema como una ampliación de la fase de Entendimiento de los Datos de la metodología CRISP-DM.

8.1.2. Segmentación.

La Segmentación tiene como objetivo la separación de datos hacia sub-grupos o clases significativos.

La segmentación se puede realizar manualmente o medianamente automática: Un analista puede empezar a tejer hipótesis a partir de subgrupos que considere relevantes a la hora de reunir las metas del negocio, tomando como punto de partida la descripción de los datos; por otro lado, existen técnicas de agrupamiento capaces de detectar patrones y estructuras antes desconocidas y escondidas en los datos.

Como ocurre con el tipo de problema Descripción y Resumen de los datos, la Segmentación también puede ser por sí mismo el objetivo principal de un proyecto de minería de datos. Aunque generalmente también corresponde a uno de los pasos hacia resolver proyectos más grandes. En ocasiones puede ser más sencillo analizar un subgrupo homogéneo que contenga las mismas características del grupo total para hacer el problema más manejable.

Para tener en cuenta: Usualmente se confunde la segmentación con los términos *Agrupamiento* (Clustering) o *Clasificación*. El último término hace mayor énfasis a la creación de clases, mientras que los demás se enfocan a la creación de modelos que predicen clases conocidas en casos desconocidos.

Algunas de las técnicas más apropiadas a la hora de abordar un problema de segmentación son:

- Técnicas de agrupamiento.
- Redes neuronales.
- Visualización.

8.1.3. Descripción de concepto.

Este problema tiene como objetivo una descripción entendible de los conceptos o clases. No se pretende desarrollar modelos completos que sean capaces de predecir con gran precisión, sino obtener información. Un ejemplo claro puede ser el interés de una compañía en conocer más sobre sus clientes a partir de los conceptos de *Lealtad y Deslealtad*. Tomando como punto de partida la descripción de estos conceptos (Clientes leales y desleales) se podría inferir qué medidas tomar para que los clientes leales lo sigan siendo, o para que los clientes leales pasen al otro grupo.

Existe una relación marcada entre la descripción de concepto y la segmentación, ya que éste último puede llevar a la enumeración de objetos que pertenecen a un concepto o clase que no tiene una descripción clara. Por lo tanto, es común trabajar primero en la segmentación para luego llevar a cabo tareas de descripción de conceptos. Aun así, existen técnicas de agrupamiento conceptual que realizan segmentación y descripción de concepto al mismo tiempo.

La descripción de concepto también se puede usar como insumo para problemas de clasificación. Algunas técnicas de clasificación producen modelos de clasificación lo suficientemente claros como para considerarlos como descripciones de concepto. La diferencia principal corresponde a que la clasificación tiene como objetivo estar completa en algún sentido, por lo que el modelo de clasificación se debe aplicar a *todos* los casos de la muestra; mientras que la descripción de concepto no requiere estar completa. Es suficiente describir partes importantes de los conceptos o clases.

Algunas de las técnicas más apropiadas a la hora de tratar con un problema de descripción de concepto son:

- Métodos de inducción de reglas.
- Agrupamiento conceptual.

Ejemplo:

Usando datos sobre los compradores de carros nuevos y usando alguna técnica de inducción de reglas, una compañía de carros podría generar reglas que describen sus clientes leales y desleales. Abajo se hace una relación de algunas reglas generadas como ejemplo:

Si	*SEXO = masculino y EDAD > 51*	*entonces CLIENTE = leal*
Si	*SEXO = femenino y EDAD >21*	*entonces CLIENTE = leal*
Si	*PROFESION = administrador y EDAD < 51*	*entonces CLIENTE = desleal*
Si	*ESTADO CIVIL = soltero y EDAD < 51*	*entonces CLIENTE = desleal*

8.1.4. Clasificación

La clasificación toma un grupo de objetos (Con ciertos atributos o rasgos) pertenecientes a diferentes clases. La etiqueta de clase es un valor discreto y conocido por cada objeto. Lo que se pretende es construir modelos de clasificación (A menudo

conocidos como *clasificadores*) que asignen la etiqueta de clase correcta a objetos desconocidos previamente y que no tengan clase alguna. Por lo tanto, la clasificación es útil a la hora de crear modelos predictivos.

La descripción de las etiquetas puede venir de mano del usuario o ser resultado de aplicar segmentación.

La clasificación es uno de los problemas más comunes en los proyectos de minería de datos y ocurren en una gama de aplicaciones amplia. Además tiene conexión con casi todos los demás tipos de problemas. Un problema de predicción se puede convertir en un problema de clasificación al combinar etiquetes de clases continuas, ya que las técnicas de combinación permiten transformar rasgos continuos en intervalos discretos. Estos intervalos discretos pueden tomarse como etiquetas de clase en vez de valores numéricos exactos. Algunas técnicas de clasificación también producen clases y descripciones de concepto entendibles.

Algunas técnicas apropiadas cuando se lidia con un problema de clasificación son:

- Análisis discriminante.
- Métodos de inducción de reglas.
- Aprendizaje de árboles de decisión.
- Redes neuronales.
- Vecino k-más-cercano.
- Razonamiento basado en casos.
- Algoritmos genéticos.

8.1.5. Predicción.

Así como ocurre con la clasificación, la predicción es uno de los tipos de problemas que más se abordan en un proyecto de minería de datos. Se diferencia de la clasificación por el tipo de atributo objetivo (Clase) al que se quiere llegar, ya que mientras que en la

clasificación se busca un valor discreto, en la predicción se busca un valor continuo. En algunas literaturas se conoce a este tipo de problema también como *Regresión*.

Algunas técnicas apropiadas para este tipo de problema son:

- Análisis de regresión.
- Árboles de regresión.
- Redes neuronales.
- Vecino k-más-cercano.
- Métodos Box-Jenkins.
- Algoritmos genéticos.

8.1.6. Análisis de dependencia.

En este tipo de problema se busca encontrar modelos que describan dependencias significantes (Asociaciones) entre datos o eventos. Esta información puede ser útil a la hora de predecir el valor de un dato, basado en la información de otros datos. Aun así, se usa análisis de dependencia en su mayoría para cuestiones de entendimiento de datos. Las dependencias halladas pueden ser estrictas o probabilísticas.

Las asociaciones son casos especiales de las dependencias y describen afinidad en datos o eventos que ocurren frecuentemente juntos. Un ejemplo sencillo y típico se da al analizar las compras en una tienda. Se puede llegar a reglas del tipo “En el 30% de las compras, se compran cervezas y maní juntos”.

Los algoritmos que detectan asociaciones son usualmente rápidos y producen resultados amplios. Lo verdaderamente interesante es seleccionar las asociaciones más relevantes.

En aplicaciones prácticas, usualmente se relacionan la segmentación y el análisis de dependencia, ya que es más recomendable aplicar análisis de dependencias en

segmentos de datos más homogéneos. En grandes grupos de datos las dependencias pueden ser irrelevantes por cómo se cubren entre sí las relaciones.

Las técnicas más usadas a la hora de analizar dependencias son:

- Análisis de correlación.
- Análisis de regresión.
- Reglas de asociación.
- Redes bayesianas.
- Programación lógica inductiva.
- Visualización.

Ejemplo:

Aplicando algoritmos de reglas de asociación a los datos sobre accesorios de carros, una compañía de carros ha descubierto que cuando se ordena un radio, se ordena una caja de cambios automática en al menos el 95& de los casos. Con esta dependencia, se decide ofrecer estos accesorios como una combinación que ayuda a reducir costos.

8.2. Árboles de decisión y Reglas de Decisión

La adquisición de conocimiento “tangible” en una época de almacenamiento masivo se convierte en un cuello de botella para la ingeniería de conocimiento. Han surgido algoritmos que extraen conocimiento a partir de datos, y que han intentado, con éxito, aligerar esta carga. Dentro de estos algoritmos, los sistemas que inducen la creación de árboles de decisión se han vuelto muy populares. Gran parte de esta popularidad se debe al fácil entendimiento del conocimiento resultante, lo que se hace atractivo no sólo para usuarios técnicos, sino para aquellos que estén interesados en el entendimiento del dominio, capacidades de clasificación, o en las reglas que pueden surgir a partir del

árbol, lo que consecuentemente podría dar lugar a un sistema de decisión basado en reglas. [21]

Los árboles de decisión se hicieron especialmente populares gracias al algoritmo ID3 de John Quinlan. Este enfoque tiene ventajas a la hora de trabajar en dominios simbólicos, pero se hacen poco prácticos en casos donde es necesario trabajar con decisiones numéricas, o en donde la decisión numérica optimiza el proceso subsecuente. [22]

En términos generales, los árboles de decisión están hechos de dos componentes principales: Un procedimiento para construir el árbol simbólico, y un procedimiento de inferencia para toma de decisiones. [23]

Por el otro lado, las reglas de decisión pueden ser descritas como una de las tantas maneras en las que se puede visualizar un árbol de decisión de manera simplificada. En términos generales, una regla de decisión es la representación de uno de los caminos posibles en un árbol de decisión desde el nodo raíz hasta uno de los nodos hoja.

8.2.1. Árboles de decisión [23]

Los algoritmos de inducción que desarrollan árboles de decisión ven la tarea de dominio como una de *Clasificación*. El marco de trabajo subyacente consiste de una colección de *Atributos* o propiedades que se usan para describir *casos* individuales, cada uno de ellos perteneciendo exactamente a un grupo (*Set*) de *Clases*.

Los atributos pueden ser continuos o discretos. El valor de un atributo continuo es siempre un número real, mientras que el valor de un atributo discreto es uno de un pequeño set de valores posibles para ese atributo. En casos de la vida real es importante tener en cuenta que un caso puede tener valores *Desconocidos* para uno o más atributos.

Un dominio discreto puede ser desordenado, parcialmente ordenado, o completamente ordenado. Cualquier tipo de ordenamiento puede ser ventajoso ya que se puede usar para acercarse a la noción de *Similitud*. Por ejemplo, si se considera el atributo *ColorDePiel*, con los valores discretos: {*Blanco, Bronceado, Negro}*. En este caso se puede decir que *Blanco* es "más similar a *Bronceado* que a *Negro*".

Un *Árbol de Decisión* puede ser una hoja identificada por un nombre de clase, o una estructura de la forma: [24]

$$
\begin{array}{ll}
C_1: & D_1 \\
C_2: & D_2 \\
 & \ldots \\
C_n: & D_n
\end{array}
$$

Donde las Ci's (Condiciones) son mutuamente exclusivas, y los Di's son por sí mismos árboles de decisión. El set de condiciones involucra sólo uno de los atributos, donde cada condición es:

$$A < T \text{ or } A > T$$

Para un atributo continuo A, y donde *T* es algún umbral; o donde:

$$A = V \text{ or } A \text{ in } \{V_i\}$$

Para un atributo discreto A, donde *V* es uno de los posibles valores y {Vi} es un subset de ellos.

Tal árbol de decisión se usa para clasificar un caso así. Si el árbol es una hoja, se determina la clase del caso para que sea nombrado por la hoja. Si el árbol es una estructura, se encuentra la única condición *Ci* que sostiene este caso y se continúa con el árbol de decisión asociado.

La siguiente figura muestra un árbol de decisión para el diagnóstico de condiciones de Hipotiroides con las siguientes clases {*Hipotiroides primaria, Hipotiroides secundaria, Hipotiroides compensada, Negativa}.* Algunos atributos como TSH y FTI son continuos y tienen valores reales, mientras que atributos como *cirugía de tiroides*, con los valores posibles {*t, f*}, son discretos. Para clasificar un caso dentro de este árbol, primero se debe preguntar si el valor de TSH es mayor a 6.05. Si el valor está debajo de este umbral se continuaría con el árbol de decisión comenzando con *T4U measured = t*, mientras que un valor por encima del umbral llevaría al árbol de decisión con nombre FTI < 64.5. En cualquier caso se continuaría de manera similar hasta llegar a una hoja.

```
TSH < 6.05:
|   T4U measured = t: negative (1918)
|   T4U measured = f:
|   |   age > 43.5: negative (58)
|   |   age < 43.5:
|   |   |   query hypothyroid = f: negative (41)
|   |   |   query hypothyroid = t: secondary hypothyroid (1)
TSH > 6.05:
|   FTI < 64.5:
|   |   thyroid surgery = f:
|   |   |   T3 < 2.3: primary hypothyroid (51)
|   |   |   T3 > 2.3:
|   |   |   |   sex = M: negative (1)
|   |   |   |   sex = F: primary hypothyroid (4)
|   |   thyroid surgery = t:
|   |   |   referral source = SVI: primary hypothyroid (1)
|   |   |   referral source = <other>: negative (2)
|   FTI > 64.5:
|   |   on thyroxine = t: negative (32)
|   |   on thyroxine = f:
|   |   |   thyroid surgery = t: negative (3)
|   |   |   thyroid surgery = f:
|   |   |   |   TT4 < 150.5: compensated hypothyroid (120)
|   |   |   |   TT4 > 150.5: negative (6)
```

Ilustración 4: Árbol de decisión para el diagnóstico de condiciones de Hipotiroides

El set de casos con clases conocidas desde el cual se induce el árbol de decisión se llama *Set de entrenamiento*. Otras colecciones de casos no vistas mientras que se desarrolla el árbol se conocen como *sets de prueba*, y se usan para evaluar el rendimiento del árbol.

8.2.1.1. Un par de ejemplos útiles [24]

ID3 y CART son los dos algoritmos de aprendizaje discriminativo más importantes que trabajan con particionamiento recursivo. Sus ideas básicas son las mismas: Particionar el espacio de muestra de una manera orientada a los datos, y representar esa partición como un árbol. En la siguiente figura se observan algunos ejemplos de árboles resultantes, construidos a partir de dos clases, ilustrados como *Negro y Blanco*. (ID3, izquierda - CART, derecha).

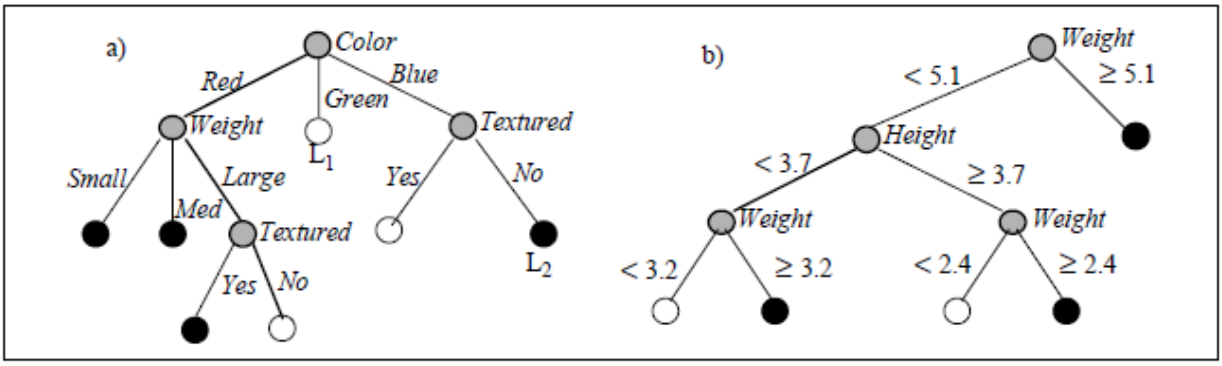

Ilustración 5: Árboles de decisión construidos con ID3 (Izquierda) y CART (Derecha)

Una propiedad importante de estos algoritmos es su propensión a minimizar el tamaño del árbol mientras se optimizan las medidas de calidad. Después, usan el mismo mecanismo de inferencia. Cuando se tiene una nueva muestra, ésta se compara contra las condiciones del árbol. Con suerte, habrá exactamente un nodo hoja cuyas condiciones (En el camino) se satisfagan. Por ejemplo, la muestra con los rasgos *[Color = Rojo] [Peso = Grande] [Con Textura = Sí]* encaja las condiciones que llevan al camino trazado por las flechas en la siguiente imagen. La clase resultante para esta muestra

corresponde a la muestra del nodo hoja al que llega, por lo que en este caso, sería de clase *Blanco*.

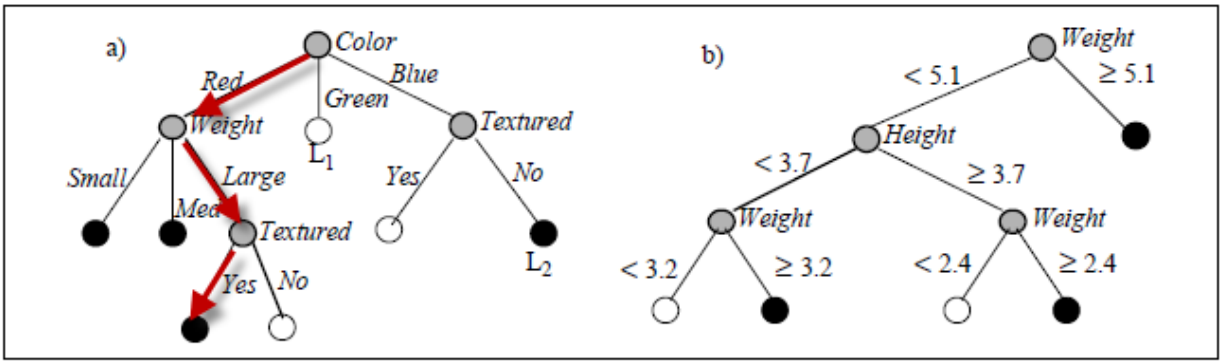

Ilustración 6: Ejemplo de árbol ID3 para las características [Color = Rojo] [Peso = Grande] [Con Textura = Sí]

ID3 asume dominios discretos con pequeñas cardinalidades, lo que representa una gran ventaja a la hora de abogar por la comprensibilidad del conocimiento. Pero para llegar a una cardinalidad así, en ocasiones es necesario un particionamiento previo. Además, cada atributo en puede proveer como máximo una condición sobre un camino dado.

CART, por otro lado, no requiere particionamiento previo, ya que las condiciones se basan en umbrales (Dominios continuos) que se calculan de manera dinámica. Esto da lugar a que un mismo atributo se pueda usar más de una vez (Con distintos umbrales) en un mismo camino. Lo que se observa en b). El número de umbrales posibles iguala el número de ejemplos de entrenamiento. Ideas así incrementan la calidad del árbol pero reducen su comprensibilidad. Una de las características especiales de CART es su capacidad de inducir nuevos rasgos a través de combinaciones lineales de rasgos ya existentes.

8.2.1.2. Reglas de Producción [24]

Esta forma de simplificación no da un árbol de decisión más pequeño, pero en cambio desarrolla un set “equivalente de *Reglas de Producción*, una representación media usada ampliamente en los sistemas expertos. El proceso tiene dos etapas: Primero se generan y “pulen” las reglas de producción individuales, y luego las reglas producidas se evalúan como una colección completa.

Cuando quiera que se use un árbol de decisión para clasificar un caso, se establece un camino entre la cima (Raíz) del árbol y una de sus hojas. Para que el caso alcance esa hoja, debe satisfacer todas las condiciones a lo largo del camino. Por ejemplo, en el árbol de la Ilustración 4, para que cualquier caso sea clasificado como *Negativo* por la última hoja del árbol (Última línea del árbol indentado) se deben satisfacer todas las condiciones:

$$TSH > 6.05$$
$$FTI > 64.5$$
$$on\ thyroxine = f$$
$$thyroid\ surgery = f$$
$$TT4 > 150.5$$

Por lo que cada hoja de un árbol de decisión se puede representar como una regla de producción que tiene la forma:

$$\textbf{if } X_1 \wedge X_2 \wedge ... \wedge X_n \textbf{ then } class\ c$$

Donde los Xi’s son condiciones como las antes mencionadas y *c* es la clase de la hoja.

El sólo reescribir un árbol como una colección de reglas de producción equivalentes no representa un gran avance. En cambio, la primera etapa examina cada regla de producción para ver si se debe generalizar quitando condiciones de su lado izquierdo. La relevancia de una condición Xi para determinar si un caso pertenece a una clase c se puede resumir a una tabla de contingencia 2 X 2.

	class c	not class c
satisfies X_i	sc	$s\overline{c}$
does not satisfy X_i	$\overline{s}c$	$\overline{sc}$

Donde sc representa el número de casos que satisfacen la condición *Xi* y al mismo tiempo pertenecen a la clase *c*. Mientras que sc representa los casos que no satisfacen la condición y no pertenecen a la clase evaluados.

	class *negative*	not class *negative*
$TSH > 6.05$	6	0
$TSH < 6.05$	154	0

En el ejemplo de arriba se observan 6 casos que satisfacen la condición TSH > 6.05, y que al mismo tiempo son negativos.

Para la última condición, TT4 > 150.5 se tiene el siguiente cuadro:

	class *negative*	not class *negative*
$TT4 > 150.5$	6	0
$TT4 < 150.5$	0	120

El proceso se repite varias veces hasta que se dé con una regla simplificada que sea relevante. Además, se calcula un *factor de certeza* para la regla simplificada, un valor útil a la hora de pasar a la segunda etapa.

En la segunda parte se observa el funcionamiento de las reglas como un grupo completo. Se adopta una estrategia simple:

Para clasificar un caso, encontrar una regla que aplique. Si hay más de una, se debe elegir aquella con el factor de certeza más alto. Si ninguna regla aplica, se debe tomar

la clase por defecto para que esta sea la clase más frecuente en el set de entrenamiento.

8.3. Análisis de Regresión

El análisis de regresión es un campo de estudio que tiene sus orígenes hace ya varios siglos, viendo sus primeros indicios a principios del siglo XIX, bajo la observación de matemáticos de renombre, como Legendre ó Gauss [25] [26], quienes buscaban determinar las órbitas de cuerpos celestes (Principalmente cometas). Es claro entonces que desde sus inicios, la regresión fue una poderosa herramienta de gran utilidad para la predicción de funciones continuas (En este caso la trayectoria de cuerpos celestes).

Por definición, el análisis de regresión es una herramienta estadística para la investigación de relaciones entre variables. Permite modelar, examinar y explorar relaciones espaciales, ayudando a explicar los factores que hay detrás de patrones espaciales observados. Lo que se busca es comprobar el efecto causal de una variable sobre otra. Para explorar problemas de ese tipo se ajustan los datos a las variables subyacentes y se emplea regresión para estimar el efecto cuantitativo y medible de una variable sobre otra(s). Además, tiene amplio uso en problemas de predicción. [27]

Algo que también se evalúa en el análisis de regresión es la Significancia Estadística de las relaciones estimadas, es decir, el grado de fiabilidad entre la relación real y la relación estimada (Qué tan cerca está la primera de la última). Se considera que entender la regresión simple es suficiente para entender los fundamentos del análisis de regresión, y para ello se usará un ejemplo práctico.

Para propósitos de ilustración, suponga que se desean identificar y cuantificar los factores que determinan las ganancias en el mercado laboral. Una revisión inicial trae a la mente una gran cantidad de factores que se asocian a las variaciones en las ganancias de una persona: Ocupación, edad, experiencia, logros educativos, motivación, habilidades innatas, o incluso el género. Por ahora, se reduce la atención a

un solo factor, que se conocerá como Educación. Cuando la variable que explica la relación es sólo una se conoce como Regresión Simple.

8.3.1. Regresión Simple.

En la realidad, cuando se intenta explicar el efecto de la educación sobre las ganancias de una persona sin prestar atención a los demás factores, se pueden crear serias dificultades estadísticas, entre los que se destaca algo conocido como Sesgo de Variables Omitidas (*OVB, Omitted-Variables Bias*). Además, se asume que la educación se puede cuantificar en un único atributo, los años de estudio. Así, se supone que un número dado de años escolares puede representar programas académicos ampliamente variados.

Al iniciar cualquier estudio de regresión es usual formular hipótesis sobre la relación entre las variables de interés, que en este caso se conocerán como Educación y Ganancias. La experiencia cotidiana sugiere que la gente mejor educada tiende a hacer más dinero. Por lo tanto, es probable que la relación causal va de la educación hacia las ganancias y no al contrario. Se tiene entonces como hipótesis tentativa que "*Entre más alto es el nivel de educación, más alto el nivel de ganancias*", con las demás cosas iguales.

Para este caso imagine que se juntan datos sobre educación y ganancias de un grupo de personas. Se conocerá *E*, como la educación en años de estudio por individuo, y se conocerá *I* a las ganancias de un individuo en dólares por año. Toda esta información se puede transmitir a un diagrama bidimensional, convencionalmente conocido como *Diagrama de Dispersión*. Cada punto en el diagrama representa a un individuo de la muestra.

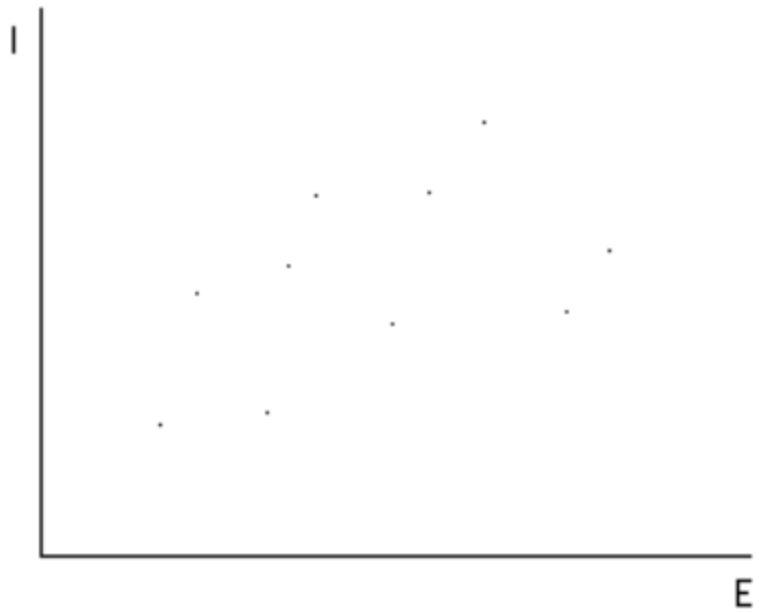

Ilustración 7: Diagrama de dispersión.

Incluso en el diagrama se sugiere que los valores más altos de *E* tienden a producir valores más altos de *I*, pero la relación no es perfecta -Al parecer el conocimiento de *E* no es suficiente para una predicción enteramente precisa de *I*. Se puede deducir entonces que el efecto de la educación sobre las ganancias difiere entre individuos, o que otros factores, distintos a la educación también influyen. El análisis de regresión acepta la última explicación sin problemas. Así, con la discusión pendiente sobre el sesgo de variables omitidas, se tiene la hipótesis de que "*Las ganancias de cada individuo están determinados por la educación y una suma de factores omitidos conocidos como Ruido*".

Para dar más forma a la hipótesis, también se supondrá que aquellas personas que trabajan y no tienen educación también ganan algo de dinero, y que la educación incrementa las ganancias a partir de este lineamiento. Se asume entonces que la educación afecta los ingresos de manera "lineal", lo que significa que cada año adicional de educación agrega la misma cantidad a los ingresos. No siempre la presunción de linealidad es esencial, ya que existen razones para pensar que la relación es *No Lineal*.

Tomando en cuenta lo anterior, se podría describir la relación Educación/Ganancia de la siguiente manera:

$$I = \alpha + \beta E + \varepsilon$$

Ecuación 1. Ec 1.

Donde:

α = Una cantidad constante (Lo que gana alguien sin educación);

β = El efecto en dólares de un año adicional de estudio en los ingresos, que se toma hipotéticamente como positivo; y
ε = El término de "ruido" que refleja los otros factores que influencian las ganancias.

La variable *I* se conoce como la variable "dependiente" o "endógena"; mientras que *E* se conoce como la variable "independiente", "explicativa" o "exógena"; α es el "término constante" y β el "coeficiente" de la variable *E*.

Además, se debe saber separar lo observable de aquello que no se puede observar. Dentro del grupo de datos hay observaciones (Valores) para *I* como para *E*. El ruido ε está compuesto de factores que no se pueden observar o que NO se han observado, así como α *y* β. La tarea principal del análisis de regresión es producir un estimado de estos dos últimos parámetros, basándose en la información del grupo de datos y también en algunas presunciones que se hacen sobre ε.

Para tener una idea de cómo se generan los parámetros estimados, tenga en cuenta que si se ignora el ruido, la ecuación planteada es la misma ecuación de una línea recta -Una línea con "intercepción" en α en el eje vertical y "pendiente" β. Llevando esto al diagrama de dispersión, la relación hipotética implica que en algún lado del diagrama se puede encontrar una recta con la ecuación:

$$I = \alpha + \beta E$$

Ecuación 2. Ec 2.

Estimar α *y* β es equivalente a estimar *dónde* se ubica esta recta.

Pero, ¿Cuál es la mejor estimación que tiene en cuenta la ubicación de esta recta? La respuesta se relaciona con lo que se asuma α del ruido ε. Si se presume que es un número usualmente negativo grande, por ejemplo, se elegiría una línea que pase sobre la mayoría de todos los puntos -La lógica es que si ε es negativo, el verdadero valor de *I* con la *Ec. 1*, será menor que el valor de *I* en la recta de la *Ec. 2*. Del mismo modo, si se asume que ε es positivo, sería más apropiado elegir una recta que pase por debajo de la mayoría de puntos. Sin embargo, el análisis de regresión asume que el ruido es en promedio igual a cero, ya que esto sugiere que la recta estimada pasa aproximadamente por el medio de los datos, con algunos datos encima y otros por debajo.

A pesar de esta reducción del dominio de estudio, sigue habiendo muchas líneas con estas características y sólo se puede elegir una. El análisis de regresión hace esto aceptando un criterio relacionado con el ruido e*stimado* de “error” para cada observación. Para ser más precisos, el “error estimado” para cada observación se define como la distancia vertical entre el valor de *I* en la Ec. 2 y el valor real de *I* en la misma observación. Sobreponiendo una línea candidata en el diagrama de dispersión, los errores estimados para cada observación se pueden ver así:

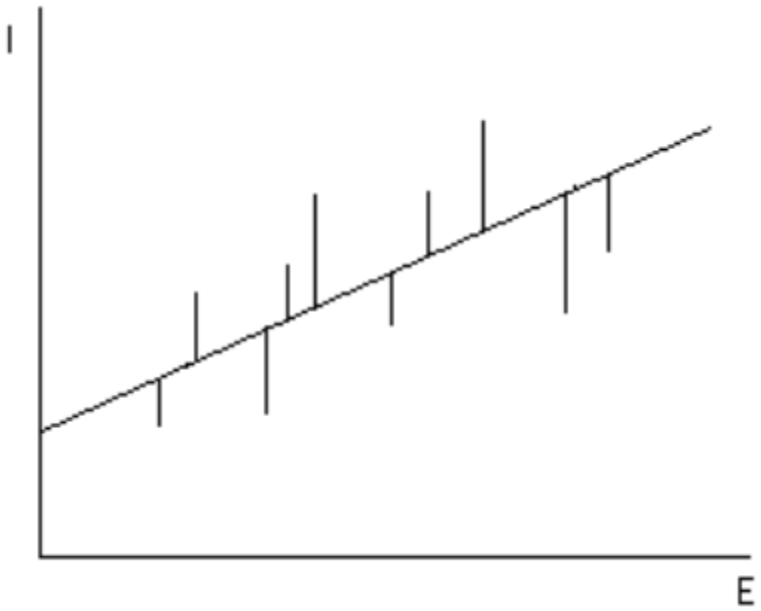

Ilustración 8: Errores en un diagrama de dispersión.

Con cada línea posible que se puede sobreponer a los datos, saldrá un set distinto de errores estimados. La regresión elige entre todas las líneas posibles aquella donde la suma de los cuadrados de los errores estimados sea el mínimo. Este criterio es conocido como Suma Mínima de Errores Cuadrados (*Minimum Sum of Squared Errors, MSSE)* o más ampliamente conocido como *Mínimos Cuadrados*. La intercepción de la línea elegida por este criterio provee el estimado de α, y la pendiente de la línea suministra el estimado de β.

Una de las ventajas a la hora de elegir mínimos cuadrados es su fácil implementación computacional, por lo que se podrían usar como datos de entrada sólo los valores de *E* y de *I.*

8.4 Agrupación o Clustering [39] [40]

El objetivo principal del clustering es obtener grupos o conjuntos de elementos, de tal forma que los elementos pertenecientes a cada grupo sean lo más similares entre ellos y diferentes a los de otros grupos como se pueda. El agrupamiento se realiza a partir

de patrones o tendencias en los datos o la distancia que hay entre los objetos. Una vez estos grupos se establezcan, pueden ser usados por otras técnicas de minería de datos.

Existe un gran número de técnicas de agrupamiento, las cuales se pueden asociar de acuerdo a la arquitectura que utilizan. Los cuatro tipos de algoritmos de agrupamiento más generales que existen son: agrupamiento jerárquico, agrupamiento basado en densidad, agrupamiento particional y agrupamiento por mixturas finitas.

8.4.1. Agrupamiento jerárquico:

Los algoritmos jerárquicos crean una descomposición de un conjunto de datos en forma jerárquica, como su nombre lo indica. Este tipo de algoritmos puede ser aglomerativo o divisivo, dependiendo de cómo se comporta su descomposición.

Los algoritmos aglomerativos son de acercamiento ascendente (bottom-up), cada dato inicia en su propio grupo y se empiezan a unir mientras suben en la jerarquía; mientras que los algoritmos divisivos son de acercamiento descendente (top-down), todos los datos inician en un solo grupo y se empiezan a hacer divisiones mientras bajan en la jerarquía.

El resultado final de este ordenamiento se representa con un diagrama en forma de árbol, el cual divide recursivamente el conjunto en más y más conjuntos cada vez más pequeños. Este tipo de ordenamiento me permite ver claramente las relaciones de agrupación entre los datos o entre los diferentes grupos.

Para que el algoritmo recursivo de agrupamiento se detenga existen algunos criterios de parada. Antes de iniciar con el algoritmo, se eligen dos datos de parada: número de grupos y distancia de umbral. Cuando en medio de las iteraciones, el número de grupos que surgen, supere al número de grupos elegidos al inicio del algoritmo, éste se

detendrá; o cuando la distancia mínima entre dos grupos, sea mayor a la distancia umbral, el algoritmo igualmente se detendrá.
Existen diferentes tipos de estrategias jerárquicas. Sin embargo, las más comunes son: el single link, Average link y Complete link. De una forma muy superficial, se tiene la siguiente definición:

- **Single Link (SL):** en cada paso se unen los dos grupos para los que sus elementos más cercanos, tienen la minima distancia.
- **Average Link (AL):** en cada paso se unen los dos grupos que tienen la mínima distancia promedio entre sus puntos.
- **Complete Link (CL):** en cada paso se unen los dos grupos en los cuales su unión tienen el diámetro mínimo o los dos grupos con la menor distancia máxima entre sus elementos.

Algunos de los algoritmos de agrupamiento más conocidos del tipo jerárquico son:

- BIRCH
- ROCK
- CHAMALEON

8.4.2. Agrupamiento basado en densidad

Este tipo de algoritmo utiliza como criterio de agrupamiento la densidad de puntos, de tal forma que los grupos que se crean muestran una alta densidad de puntos en su interior, mientras que entre ellos aparecen zonas de densidad baja.

Estos algoritmos usan diversas técnicas para determinar los grupos las, dentro de las cuales pueden ser la implementación de grafos, basadas en histogramas, kernels, aplicando la regla k-NN, empleando los conceptos de punto central, borde o ruido.

Los algoritmos basados en densidad más comunes son:

- DBSCAN: Density Based Spatial Clustering of Application whit Noise.
- OPTICS: Ordering Points To Identify the Clustering Structure.
- DENCLUDE: Density-based Clustering.
- CLIQUE: Clustering in Quest.
- SNN: Shared Nearest Neighbor, density-based clustering.

El primer algoritmo que empleó este tipo de agrupamiento para dividir el conjunto de datos fue el DBSCAN, en este existen los conceptos: punto central, punto borde y punto ruido; los que son empleados para determinar los diferentes clusters.

8.4.2.1. Algoritmo DBSCAN

Este algoritmo da inicio seleccionando un punto arbitrario p, si p es un punto central (puntos que tienen en su vecindad una cantidad de puntos mayor o igual que un umbral especificado), se comienza a construir un grupo en el cual se van a ubicar todos los objetos alcanzables desde p. Si p no es un punto central se pasa a otro objeto del conjunto de datos. Este proceso se repite hasta que son procesados todos los puntos. Los puntos que quedan fuera de los grupos formados son los llamados puntos ruido, y los que no son puntos ruido ni puntos centrales se les denomina puntos borde. De esta forma DBSCAN construye grupos en los que sus puntos son o puntos centrales o puntos borde, un grupo puede tener más de un punto central.

En la siguiente imagen se puede observar una región densa de puntos, sin agrupar en la izquierda y agrupadas mediante DBSCAN en la derecha:

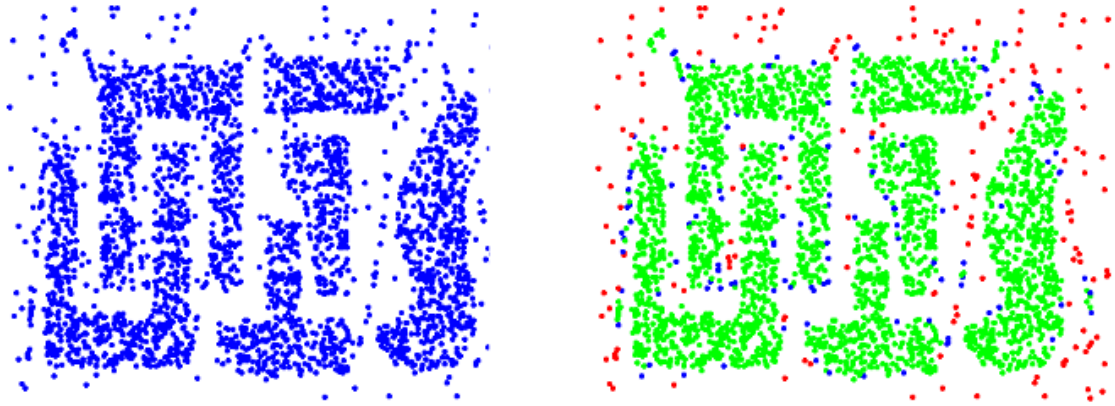

Ilustración 9: Agrupamiento DBSCAN.

En la región que está agrupada, los puntos centrales (Cluster) se ven identificados con el color verde, los puntos borde con el color azul y los puntos ruido con el color rojo. Vale la pena indicar que este algoritmo tiene una eficiencia de O(n log n).

8.4.3. Agrupamiento particional

Los agrupamientos de tipo particional organizan los objetos dentro de un número k de grupos (determinado previamente) de tal forma que se minimizada la desviación total de cada objeto desde el centro de su grupo o desde una distribución de grupos.

Este tipo de agrupamiento cuenta con un inconveniente muy notable, y es que el número de grupos que se van a usar se debe elegir desde antes de empezar el algoritmo. Así que posiblemente el número de clusters que se tengan al final no sea el óptimo para el agrupamiento necesario.

Los algoritmos más importantes de tipo particional son el K-Means, el CURE y el CLARANS; siendo el primero el más conocido y del cual se dará una breve explicación.

8.4.3.1. Algoritmo K-Means

El algoritmo más conocido del tipo particional es también uno de los más simples en su categoría. Sigue un algoritmo simple para dividir una base de datos en k grupos, los cuales se fijan con anterioridad.

Este algoritmo funciona con la idea principal de definir un centroide para cada uno de los grupos que se haya decidido tener. Luego de haber definido los centroides los demás puntos se asignan al cluster cuyo centroide esté más cerca, esto en base a cualquier métrica de distancia que se haya decidido utilizar. Luego se recalculan los centroides en función a la asignación de puntos, y se repite el proceso iterativamente hasta que los centroides dejen de cambiar.

El problema con este tipo de algoritmo radica en la selección a priori de los centroides. Para tratar esta desventaja se recomienda realizar varias ejecuciones de algunos pasos del algoritmo con diferentes centroides iniciales y comparar resultados. O se puede realizar primero un algoritmo de tipo jerárquico sobre una muestra del grupo, esto para elegir unos buenos centroides iniciales para el algoritmo.

K-Means es un algoritmo tan extenso y estudiado que posee unas variantes dentro de sí mismo: como lo es GRASP, el cual es usado para optimizar la selección de centroides iniciales; K-Modes, que utiliza modas en lugar d emedias para poder trabajar con datos tipo categórico; o K-Medoids, en el cual se utilizan medianas en vez de medias para limitar la influencia de los grupos perdidos o outliers.

8.4.4. Agrupamiento por mixturas finitas

Este método de agrupamiento es diferente a los anteriores, y es que a diferencia de los demás, éste es un tipo de agrupamiento paramétrico. El agrupamiento por mixturas finitas es una herramienta poderosa para modelar densidades de probabilidad de conjuntos de datos univariados y multivariados. Modelan observaciones las cuales se asume que han sido producidas por un conjunto de fuentes aleatorias alternativas e

infieren los parámetros de estas fuentes para identificar qué fuente produjo cada observación, lo que lleva a un agrupamiento del conjunto de observaciones.

Se tiene una variable Y = [Y1, Y2,…, Yd] ^{T} una variable aleatoria d-dimensional con y = [y1, y2,…, yd] ^{T} representando un resultado particular de Y. Se puede afirmar que Y sigue una distribución de mixtura finita con k componentes si su función de densidad de probabilidad se puede escribir por:

$$p(y/\theta) = \sum_{m=1}^{k} \alpha_m \; p(y/\theta_m)$$

Donde α1, α2,…, αk son probabilidades mezclantes que nos indican el grado de importancia de cada uno de los k modos, cada Θm es el vector de parámetros que define la m-ésima componente, Θ = {Θ1,…, Θk, α1,…, αk} es el conjunto completo de parámetros necesarios para especificar la mixtura, las αm deben satisfacer la siguiente ecuación:

$$\alpha_m \geq 0 \;\text{ para todo m=1,…, k y }\; \sum_{m=1}^{k} \alpha_m = 1$$

La opción usual para obtener los estimados de los parámetros es el algoritmo Expectation-Maximization (EM), que produce una secuencia de estimados de Θ aplicando alternativamente dos pasos (E-paso y M-paso) hasta obtener un máximo local de: log p (Y/Θ), siendo Y un conjunto de n muestras independientes e idénticamente distribuidas.

CAPITULO V.

9. Conclusiones

- La información debe ser vista como recurso general dentro de las compañías y no como propiedad de alguien en particular en la organización. Los productos que son ricos en característica intelectual aumentan su valor con la frecuencia de uso.

- Cuando se cuenta con una gran cantidad de datos pero estos no son usados en ningún proceso de la compañía se vuelven un costo para la misma. Para mantener los datos se necesitan elementos de hardware que los almacene, software que los administre y recursos de sistemas que cuiden de ellos, pero nadie los usa.

- La implementación de una herramienta de BI en compañía de un sistema ERP, potencializa el valor agregado que generan ambas herramientas por sí mismas, ya que con un sistema ERP se cuenta con un suministro de información lo suficientemente amplio y confiable como para mantener bien alimentado una herramienta de BI que apoyará la toma de decisiones dentro de la empresa.

- Una solución de BI brinda acompañamiento con herramientas tecnológicas sobre todo el proceso de transformación de los datos ordinarios de la empresa en inteligencia para la misma, garantizando confiabilidad, integridad, velocidad y seguridad en todo el ciclo del conocimiento, desde la captura de los datos hasta la visualización de la información en tableros de control o herramientas de reporting.

- Contar con herramientas de alto poder de procesamiento a la hora de darle buen uso a la información es posible gracias a conceptos como el Data Warehousing

(DW), que abarcan disciplinas completas de las cuales se podrían hacer estudios extensos. Hablar de DW no se refiere sólo a tener un repositorio donde se puedan almacenar los datos operativos de una empresa. En cambio, hace referencia a todo un conjunto de tecnologías que comprende desde la recolección de los datos de distintas fuentes con distintos formatos; pasando por un proceso de transformación que resulta en datos convergentes; y finalizando en operaciones que permiten extraer la información y analizarla desde puntos de vista del más alto nivel. Esta facilidad provista por DW permite a gerentes y administradores tener información precisa a la mano y a un nivel de entendimiento propio de su cargo.

- En la actualidad hay grupos de investigación dedicados a perfilar el campo del DW hacia las distintas áreas de negocio aprovechando tecnologías existentes como lo son la Web y la reciente y en pleno furor Nube. Para el usuario gerencial podría ser transparente la infraestructura detrás de su herramienta de DW, mientras ésta no resulte en una relación costo/beneficio perjudicial. Por el contrario, para el usuario dedicado al diseño e implementación de una herramienta de DW contar con mejoras en el procesamiento y optimización en los procesos principales puede ser trascendental. Por consiguiente, mantenerse al tanto de las últimas corrientes es una práctica recomendada.

- Los conceptos de minería de datos aquí relacionados son quizá la esencia de este documento, pues pretenden explicar, en un término medio entre lo breve y lo profundo, la teoría que subyace a la resolución de problemas de toma de decisiones. Pretende ser una guía de referencia a la hora de buscar conceptos que, aunque no son nuevos, pueden encontrarse dispersos en diversa literatura.

- Estudiar CRISP-DM es una manera óptima para entender el corazón de la minería de datos. Aunque en este documento se explica de manera breve, se recomienda su posterior lectura, sobre todo si se piensa implementar un

proyecto de minería de datos y se quiere contar con un marco de trabajo confiable y recomendado ampliamente en la comunidad de minería de datos.

- Dentro de las cuestiones principales de la minería de datos, la elección de la técnica o algoritmo adecuado para el problema puede ser un dolor de cabeza. Saber de entrada cuáles son los objetivos del negocio, y tener una representación de éstos en términos de objetivos de minería de datos, es un primer paso. Es útil contar con la experiencia de aquellos que han enfrentado situaciones similares y para eso las comunidades virtuales juegan un papel importante.

- Finalmente, consideramos que este documento representará una nueva opción gracias a que converge temas enmarcados dentro de la Ingeniería de Conocimiento y que en principio pueden aparecer dispersos en diversas literaturas. El lenguaje usado es una combinación de lenguaje técnico, propio de las fuentes, con lenguaje cotidiano, ya que está dirigido a usuarios de alto nivel (Usualmente gerenciales) o de bajo nivel (Un analista de datos)

10. Bibliografía

[1] "Minería de Datos - Trabajo de Adscripción"; Sofía J. Vallejos; 2006; http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf

[2] "**CRISP-DM 1.0 Step-by-step data mining guide"; Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer & Rüdiger Wirth; 2000;** http://www.nancygrady.info/CRISP-DM.pdf

[3] "KDnuggets - Data Mining Community Top Resource"; [En línea]; http://www.kdnuggets.com/

[4] "**From DSS to DSP: A taxonomic retrospective"; Arun Sen; Magazine "Communications of the ACM - Volume 41 Issue 5es, May 1998 / Pages 206 - 216**

[5] "A Data Mining & Knowledge Discovery Process Model"; Óscar Marbán, Gonzalo Mariscal, Javier Segovia; Data Mining and Knowledge Discovery in Real Life Applications; 2009

[6] "A critical analysis of Decision Support Systems research"; Arnott, D. and Pervan, G.; Journal of Information Technology, 20, 2, June, 2005, pp67-87

[7] "A Very Short History of Data Science"; [En línea]; Gil Press; 2013; http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

[8] "What main methodology are you using for data mining? (Jul 2002)"; [En línea]; KD Nuggets; 2002; http://www.kdnuggets.com/polls/2002/methodology.htm

[9] "Data Mining Methodology (Apr 2004)"; [En línea]; KD Nuggets; 2004; http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm

[10] "Data Mining Methodology (Aug 2007)"; [En línea]; KD Nuggets; 2007; http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

[11] "What is Alter's DSS Taxonomy?"; [En línea]; Dan Power; 2008; http://dssresources.com/faq/index.php?action=artikel&id=167

[12] "Do people really understand data warehousing?"; [En línea]; Rick Sherman; 2008;
http://searchmanufacturingerp.techtarget.com/news/1327878/Do-people-really-understand-data-warehousing?bucket=NEWS

[13] "About Bill"; [En línea]; 2007; http://www.inmoncif.com/about/

[14] "Building the Data Warehouse"; William H. Inmon; 2000

[15] "About Us - SearchBusinessIntelligence.IN - About TechTarget, Inc."; [En línea]; http://searchbusinessintelligence.techtarget.in/about

[16] "Guía metodológica para el estudio y utilización de la plataforma de inteligencia de negocios Oracle Business Intelligence Standard Edition One"; Reinero Moreno Ocampo; 2012

[17] "Three ETL process Essentials"; [En línea]; Derick Jose; 2012;
http://searchbusinessintelligence.techtarget.in/tip/Three-ETL-process-essentials

[18] "OLAP (online analytical processing"; [En línea]; whatis.com; 2001; http://searchcrm.techtarget.com/tip/OLAP-online-analytical-processing

[19] "From Data Mining to Knowledge Discovery in Databases"; Usama Fayyad, Gregory Piatetsky-Shapiro & Padhraic Smyth; AI Magazine , Volume 17, Number 3; 1996; http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131

[20] "KDnuggets - Polls - Algorithms for data analysis / data mining"; [En línea]; 2011; http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html

[21] "Understanding the Nature of Learning"; R.S. Michalski, J. Carbonell & T. Mitchell; In Machine Learning: An Artificial Intelligence Approach (eds.), Vol, II, pp. 3-26. Morgan Kaufmann, 1986

[22] "Induction on Decision Trees"; J.R. Quinlan; Machine Learning, Vol. 1; pp. 81-106; 1986

[23] "Fuzzy Decision Trees: Issues and Methods"; Cezary Z. Janikow; 1996

[24] "Simplifying Decision Trees"; J.R. Quinlan; 1986

[25] "Nouvelles méthodes pour la détermination des orbites des cométes"; Adrien Marie Legendre; 1805

[26] "Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum"; Carl Friedrich Gauss; 1809

[27] "An Introduction to Regression Analysis"; Alan O. Sykes; The Inaugural Coase Lecture; 1992

[28] "Industrial intelligence - a business intelligence-based approach to enhance manufacturing engineering in industrial companies"; Heiner Lasi; 8th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 2013

[29] "A Decision Support System Based on GIS for Flood Prevention Of Quanzhou City"; Yupeng Huang, Wenchen Lin, Haijiang Zheng; Fifth International Conference on Intelligent Human-Machine Systems and Cybernetics, 2013.

[30] "A proposed model for data warehouse ETL processes"; Shaker H. Ali El-Sappagh, Addeltawab M. Ahmed Hendawi, Ali Hamed El Bastawissy; Journal of King Saud University - Computer and Information Sciences 23, 2011

[31] "Data mining - past, present and future - a typical survey on data streams"; M.S.B. PhridviRaja, C.V. GuruRaob; Procedia Technology 12, 2014

[32] "Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods"; Hui Li, Jie Sun, Jian Wu; Expert systems with applications Journal; Volume 37, Issue 8, August 2010, Pages 5895-5904

[33] "The CART decision tree for mining data streams"; Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, Piotr Duda; Information Sciences; Volume 266, May 2014, Pages 1-15

[34] "Hierarchy classification for Data Warehouse: A Survey"; Kanika Talwar, Anjana Gosain; Procedia Technology; Volume 6, 2012, Pages 460-468

[35] "Integration of Business Intelligence and Enterprise Resource Planning within Organizations"; Muhmadd I. Nofal, Zawiyah M. Yusof; Procedia Technology; Volume 11, 2013, Pages 658-665

[36] "Social Business Intelligence: A New Perspective for Decision Makers"; Mihaela Muntean, Liviu Gabriel Cabâu, Vlad Rînciog; Procedia - Social and Behavioral Sciences; Volume 124, March 2014, Pages 562-567

[37] "Computing agents for decision support systems"; D. Krzywicki, L. Faber, A. Byrski, M. Kisiel-Dorohinicki; Future Generation Computer Systems; Volume 37, July 2014, Pages 390-400

[38] "A decision-support system for the design and management of warehousing systems"; Riccardo Accorsi, Riccardo Manzini, Fausto Maranesi; Computers in Industry; Volume 65, Issue 1, January 2014, Pages 175-186

[39] "Algoritmos de agrupamiento"; D. Pascual, Departamento de Computación, Universidad de Oriente, Santiago de Cuba, Cuba; F. Pla, S. Sánchez, Departamento de Lenguajes y Sistemas Informáticos, Universidad Jaume I, Castellón, España; 2007

[40] "Modelo clustering para el análisis en la ejecución de procesos de negocio"; Surelys Pérez Jiménez, Departamento Ingeniería Industrial, Instituto Superior Politécnico José Antonio Echeverría; Joan Jaime Puldón, Facultad Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría Rafael A; Espín Andrade, Centro de Estudios de Técnicas de Dirección (CETDIR), Instituto Superior Politécnico José Antonio Echeverría; 2012