UNIVERSIDAD DE CIENCIAS EMPRESARIALES Y SOCIALES

MAESTRÍA EN ADMINISTRACIÓN Y MARKETING ESTRATÉGICO

>> TESIS DE GRADUACIÓN

MINERÍA Y ANÁLISIS DE DATOS APLICADOS EN UN PROGRAMA DE FIDELIZACIÓN DE CLIENTES MULTIMARCAS

Tutor: Ing. María del Rosario Bruera

Alumna: Lic. Marcela Viviana Sinisgalli

2002<<

Índice y Contenidos

Minería y Análisis de Datos aplicados en un Programa de Fidelización de Clientes Multimarca

OBJETIVOS

Los proyectos de Data Mining requieren para su éxito de un importante soporte metodológico y capacitación de los usuarios en las técnicas de análisis y la implementación de la tecnología adecuada.

El éxito de un proyecto de Data Mining depende – más que de las herramientas y recursos de hardware y software – de que se formulen las adecuadas **preguntas de negocio** y se seleccionen los correctos **caminos de análisis y gestión de los datos**. Estos procesos aún no han podido automatizarse y por lo tanto de ningún modo están resueltos por la "inteligencia" incorporada a las herramientas y deben ser realizados "manualmente" con el imprescindible aporte del criterio y conocimiento del negocio de los analistas de datos.

El presente trabajo tiene dos objetivos centrales:

1) Por un lado, presentar los fundamentos del Data Mining y sus posibilidades metodológicas, aplicados a un caso de negocio real como lo es un Programa de Fidelización de Clientes.

   Con ello se pretenderá demostrar, sobre la base de la metodología desarrollada, de qué manera se lleva a cabo un proyecto de Minería de Datos, cuáles pueden ser los aportes concretos al negocio y de qué manera dichos aportes sirven de piedra angular para el desarrollo de las estrategias comerciales y de marketing.

   Sin lugar a dudas, un Programa de Fidelización de Clientes es una fuente de datos muy valiosa y de gran potencial para la aplicación de técnicas y procesos exploratorios.

   De esta forma tratará de demostrarse que el Data Mining no es un producto ni un sistema que puede adquirirse e implementarse, sino más bien es un proceso y una metodología de análisis continuo, en el que intervienen todas las áreas de la empresa, y cuyos resultados sirven de soporte para la toma de decisiones estratégicas.

2) Por otro lado, se abordará sobre la integración que existe entre el Data Mining y otras técnicas de exploración aplicada como lo es la Investigación de Mercados, en el contexto del Programa de Fidelización.

   Con ello pretenderá demostrarse que la integración de ambas técnicas –Data Mining e Investigación de Mercados- permite arribar al verdadero conocimiento de los clientes fundiendo la vista transaccional u operativa con el comportamiento actitudinal de los mismos.

   La información resultante de este proceso permitirá desarrollar las estrategias de relacionamiento con los clientes, en el marco del Control de Gestión requerido para el Programa de Fidelización.

Minería y Análisis de Datos aplicados en un Programa de Fidelización de Clientes Multimarca

HIPÓTESIS DE TRABAJO

DEMOSTRAR DE QUÉ MANERA LOS MODELOS TEÓRICOS DE ANÁLISIS Y "DATA MINING", LA INVESTIGACIÓN DE MERCADOS Y LAS TECNOLOGÍAS DE INFORMACIÓN SE INTEGRAN EN LA REALIDAD DEL NEGOCIO,

PARA ORIENTAR LA ESTRATEGIA DE RELACIONAMIENTO CON LOS CLIENTES Y EL CONTROL DE GESTIÓN EN UN PROGRAMA DE FIDELIZACIÓN DE CLIENTES MULTIMARCA.

Como se ha desarrollado en los Objetivos del Trabajo, pretenderá demostrarse de qué manera el Data Mining y las demás técnicas exploratorias sirven a los negocios para el desarrollo de las estrategias comerciales.

Considerar estos aspectos como punto de partida resulta sumamente importante, ya que de ésta forma, queda garantizado el máximo y verdadero conocimiento de los Clientes y el planteamiento de estrategias comerciales sobre una base de información analítica, consistente y fehacientemente demostrable.

ESTRUCTURA DEL TRABAJO

El presente trabajo se estructura en los siguientes capítulos con el objeto de facilitar al lector la comprensión de todos los aspectos que influyen en el tema.

- **Introducción**

- **Capítulo I – El Ciclo de Relacionamiento con los Clientes.**

  - Comprende un desarrollo del Marketing de Relaciones y su enfoque en técnicas analíticas como respaldo de las estrategias de relacionamiento con los clientes.

- **Capítulo II – Gestión Estratégica de la Información para la búsqueda de Conocimiento.**

  - Este Capítulo constituye un soporte imprescindible para la comprensión de todos los aspectos desarrollados en la Aplicación Práctica:

    - Brinda un marco conceptual y metodológico sobre los alcances de las técnicas de búsqueda de conocimiento -Data Mining- en el contexto de negocios.
    - Incluye el desarrollo de las principales técnicas de exploración de los datos y su integración en la Gestión de Relacionamiento con los Clientes.
    - Incluye también un apartado con el desarrollo de una arquitectura típica de Data Warehouse que introduce al lector en el concepto de almacenamiento multidimensional.

- **Capítulo III – Aplicación del Data Mining en un Programa de Fidelización Multimarca.**

  - Desarrolla una aplicación real de minería de datos en el contexto del Programa de Fidelización de Clientes Multimarca Travelpass. La aplicación muestra de qué manera los modelos de análisis, la investigación de mercados y la integración de técnicas sirven de soporte para el la Estrategia de Relacionamiento de Clientes y Control de Gestión.

- **Conclusiones**

- **Citas Bibliográficas**

Minería y Análisis de Datos aplicados en un Programa de Fidelización de Clientes Multimarca

INTRODUCCIÓN

Las organizaciones se enfrentan a nuevos desafíos que requieren el aprovechamiento inteligente de su propia información para discernir la evolución de los negocios.

La cantidad de información generada por las organizaciones sólo tiene una tendencia y es aumentar. A medida que las actividades comerciales y las transacciones electrónicas se han hecho más comunes en Internet y las tendencias de relacionamiento con el cliente han propiciado la generación de **vínculos directos y personalizados** con los mismos, la generación de información es aún mayor.

Las organizaciones de hoy en día deben confrontar las tareas de gestión con el **análisis de datos**. La gestión de los datos se está viendo beneficiada del concepto de almacenamiento de datos, que integra datos en información apropiada para las aplicaciones analíticas. El **análisis de datos** se ha visto beneficiado por la existencia de nuevas técnicas que permiten encontrar rápidamente las respuestas a preguntas de negocio tales como:

¿Quiénes son mis mejores clientes?

¿Cuánto debo invertir en retener a un cliente?

¿Qué productos se venden en forma conjunta?

¿Cuáles de mis clientes están a punto de abandonar mi negocio?

¿Quiénes responderán a una campaña a través de un mailing?

Las nuevas tecnologías de análisis de datos – o búsqueda de conocimiento - resultan entonces una ventaja evidente para aquellos que las adoptan e integran a sus procesos comerciales.

Con información apropiada se disminuyen costos, se mejora la relación con el cliente, se orienta el diseño de los productos, se optimizan los recursos de las empresas.

Minería y Análisis de Datos
aplicados en un Programa de Fidelización de Clientes Multimarca

CAPÍTULO I
LA RELACIÓN CON EL CLIENTE

Marketing De Relaciones

La elaboración de estrategias comerciales ha cambiado mucho desde los principios del nuevo milenio. Los estrategas se enfrentan ante un tipo de cliente con características absolutamente diferentes del que conocían hasta hace tan sólo diez años. Éste se ha convertido en un blanco móvil y cambiante. El comportamiento que observa es difuso y difícil de rastrear. Ostenta una gran volatilidad en los usos y costumbre, lo que exige un especial dinamismo a la hora de intentar satisfacer sus deseos. Lo caracteriza un vínculo cada vez más labil con la marca, situación que lleva a invertir grandes esfuerzos dedicados a incrementar la fidelidad. El grado de exigencia respecto de los atributos funcionales y simbólicos que debería tener el producto es cada vez más grande y las diferencias son fácilmente olvidadas. El extraño objeto del deseo deja de serlo, en forma cada vez más rápida. El conocimiento de los productos y servicios que está dispuesto a adquirir es mucho mayor, facilitado por el desarrollo de los medios de comunicación y últimamente por la transparencia y la rapidez de acceso a diferentes alternativas que le otorga la Red. A su vez, el comprador se ha ido enmascarando detrás de un complejo proceso, dentro del cual resulta muy difícil identificar quien es finalmente el decididor, lo que complica seriamente la definición del blanco de mercado a atacar.

A lo expuesto deben sumársele diversos cambios acaecidos en el escenario social. La integración al trabajo fuera del hogar de la mujer y el poder de decisión adquirido por los niños a la hora de comprar, han exigido la elaboración de estrategias muy diferentes a las utilizadas tradicionalmente. Por su lado, los deseos han evolucionado a un ritmo antes desconocido, como producto de una demanda de confort y calidad de vida cada vez mayor. Los cambios han penetrado hasta las raíces más profundas del núcleo familiar, enfrentando a las organizaciones ante la transición de una familia celular a una nuclear, con redes de decisión muy diferentes. La diversión ostenta una tendencia a dejar de ser colectiva para convertirse cada vez más individual y el grado de información ha crecido en magnitudes espectaculares. Los precios se ven acompañados por una comoditización de la calidad, que en ciertos casos ha desplazado el centro de la decisión de compra a variables mucho más sofisticadas y difíciles de conducir.

En definitiva, se está ante la presencia de un consumidor muy diferente, más experimentado, más capacitado, mejor informado y mucho más exigente, que asume como masa crítica un precio ajustado y una excelencia en la calidad. **Es un cliente que exige una atención mucho más personalizada y directa** que el consumidor tratado bajo los paradigmas del marketing masivo.

La desaparición del mercado de masas y la aparición de una segmentación cada vez mayor lleva a las empresas a reflexionar seriamente sobre un cambio de paradigma comercial. En efecto, un largo camino se ha recorrido, desde la época del marketing concentrado en la venta, hasta el actual que suele denominarse de relaciones, pasando por el reactivo, consistente en estimular las sugerencias y el proactivo, orientado hacia el contacto continuo para escuchar al consumidor. Se trata de un cambio sustancial, tendiente a establecer una asociación con el cliente que permita colaborar con él, para optimizar su esfuerzo, otorgarle más beneficios y mejorar el desempeño. Es una migración radical de un Marketing de transacciones a un Marketing de Relaciones. De una Ingeniería de Productos a una Ingeniería de Relaciones.

El éxito de las organizaciones de principios del tercer milenio, está en cumplir con dos requisitos que se convierten en condición necesaria:

- Colocar al cliente en el centro de la estrategia, escuchándolo y satisfaciéndolo
- Actuar con rapidez y sorpresa buscando continuamente diferencias.

La relación con el cliente es el eje del nuevo paradigma. Poner los ojos en la creación, mantenimiento y mejora de la relación con el cliente, implica dejar atrás conceptos como participación del mercado o volumen de ventas para cambiarlos por otros mucho más vitales para la vida de la organización comercial como tasa de retención, costo de obtención de nuevos clientes y utilidad unitaria y vida media por cliente. En definitivas, se trata de concentrar el análisis en el valor del cliente: Es la búsqueda de la participación en el cliente, en lugar del

deseo desmedido de lograr a cualquier costo participación en el mercado. Una progresiva participación en el cliente llevará en el largo plazo a aumentar las ventas totales e incrementar consecuentemente la participación en el mercado. La nueva concepción no sugiere dejar de apropiarse de los beneficios de las producciones en masa y el aprovechamiento de las economías de escala, sino que intenta diseñar una red de servicios para el cliente cada vez mejor y entablar un diálogo cada vez más profundo con los consumidores. Busca alcanzar un grado de calidad en la relación con el cliente que vaya creciendo a lo largo del tiempo de manera que se plasme en un mayor índice de fidelidad y en un creciente volumen de compra.

Una estrategia como la comentada exige un **conocimiento cada vez más profundo de los clientes en forma individual**. Ante todo debe ponerse el foco en el potencial comercial que el mismo es capaz de generar. La mejor forma de acceder a esa información es analizando su comportamiento, oscultando en sus preferencias e investigando acerca de sus compras. Los medios actuales que provee la tecnología permiten entrar en el mundo de **la interactividad** que le promete resultados excelentes al marketing relacional para conocer acerca de los deseos del cliente y poder satisfacerlos. Muchas organizaciones premian a los clientes que les suministran información, de manera de ir formando de la manera más rápida posible una confiable fuente de datos. Cuanto más información se tiene de los clientes, más profunda y fructífera será la relación que con ellos se puede entablar y menor la posibilidad de que se aleje de la transacción.

Existen algunas preguntas elementales que suelen hacerse con el objetivo de ir conociendo a los clientes que conforman el patrimonio de la organización y cumplimentando el **primer paso** de la implantación de un sistema de Marketing Relacional:

- **Quién es el cliente.**
- **Qué características tiene.**
- **Con qué periodicidad compra.**
- **Cuándo compró la última vez.**
- **Cuánto compra promedio cada vez.**
- **Qué le compra a los competidores.**
- **Qué deseos tiene el cliente.**

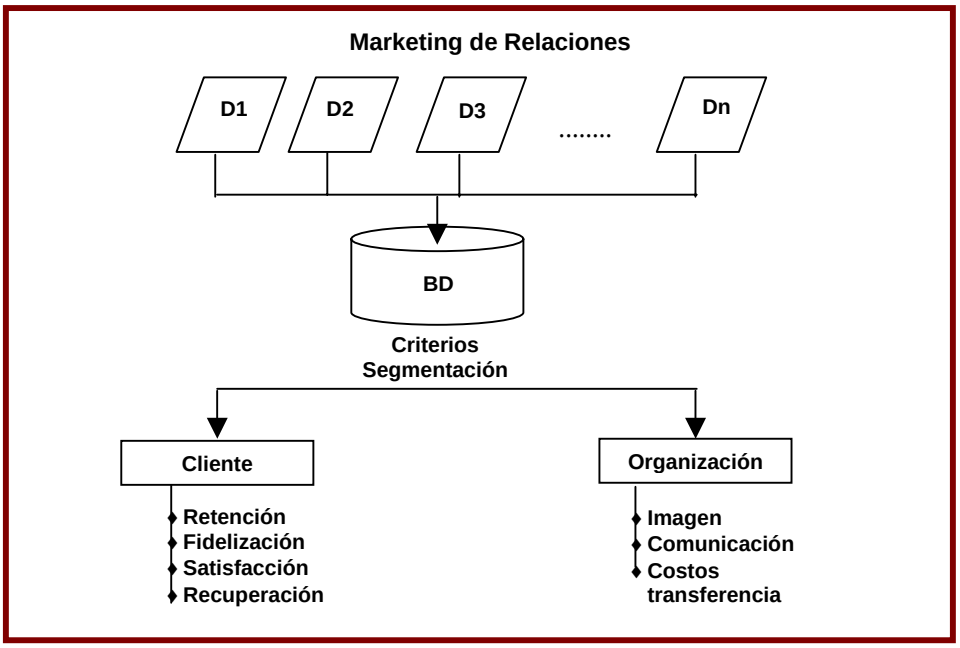
Esta información permitirá diferenciar a la cartera de clientes, de manera de poder clasificarlos por su valor vitalicio y estratégico, de forma de establecer, basándose en el primer índice, el segmento considerado estratégicamente clave. Un análisis pormenorizado del segundo permitirá cimentar las bases de la plataforma de oportunidades de crecimiento.

En un **segundo paso** se intentará adaptar las acciones de marketing de la organización, los productos y sobre todo los servicios, al perfil y deseo de los clientes. Para un **tercer paso** ajustar los medios de llegada y comunicación con el consumidor individual, de manera de establecer el diálogo más efectivo posible. Finalmente, se convierte en indispensable crear un sistema de seguimiento y control de la relación a través del tiempo, de forma de ir mejorando progresivamente el valor del cliente para la empresa.

Los beneficios que se espera alcanzar en forma directa son diversos y consisten en lograr una mayor retención, incrementar la fidelidad, diferenciarse continuamente, participar cada vez más en el consumo del cliente y seducir a clientes de los competidores. Indirectamente se logrará un ahorro de costos mejorando el índice de caída, contribuyendo al proceso de innovación y desarrollo de nuevos productos y logrando un cliente más satisfecho que difundirá las aptitudes de la compañía en el mercado.

El sistema incluye diferentes fuentes de ingreso de datos. Los clientes pueden contestar a muchas de las preguntas anteriormente enumeradas a través del call center, mediante una

encuesta de satisfacción al cliente, en oportunidad de realizar una compra, cuando tengan que presentar una queja o en situaciones en que tienen dudas acerca de las prestaciones del producto, tan sólo como para nombrar alguna de las interacciones posibles. Los datos provistos irán completando progresivamente una base de datos unificada, que será tratada continuamente para no contar con información redundante, superflua o innecesaria. Como lo muestra la **Figura-Marketing de Relaciones**, la gran utilidad que tendrá la base de datos disponible será la de permitir segmentar a los clientes según diversos criterios, entre los cuales se destaca el del valor que ellos tienen para la organización, de manera de estructurar la mejor campaña para cada uno, según el grado de interés de la organización y los beneficios por ellos buscados.



La **segmentación** obtenida permitirá estructurar acciones valiosas de marketing en áreas típicas del nuevo marketing dinámico de relaciones, tales como las vinculadas a la revitalización del vínculo con el cliente, el fortalecimiento de la retención, la generación de ámbitos y marcos de contención e involucramiento, de manera de enriquecer la fidelidad hacia la marca, la creación de planes tendientes a mejorar los procesos de forma de satisfacer cada vez más los deseos de los clientes y el análisis del abandono como fuente de creación de planes de recuperación de clientes.

Los réditos, adicionalmente, deberían ser importantes a la hora de trabajar en todas aquellas actividades vinculadas al fortalecimiento de la marca, creación de barreras importantes hacia la migración de competidores y eficiencia en los medios y características del mensaje a utilizar para comunicar. Toda vez que se posee un conocimiento mucho más pormenorizado y afinado del perfil y expectativa de los clientes.

Dos divisiones del Marketing

La relación con el cliente es el eje central sobre el que se construye el nuevo marketing de relaciones que se propone como concepto subyacente en una estrategia dinámica orientada a alterar continuamente el equilibrio del tablero competitivo mediante perturbaciones rápidas y sorpresivas, basadas en un conocimiento más profundo de los deseos y expectativas. Un marketing que intenta crear, profundizar y optimizar la interacción continua con el cliente. Enfoque que dista sustancialmente del propuesto por el marketing tradicional orientado a trabajar sobre el producto para lograr ventajas diferenciales sustentables y defendibles a ultranza. Sin embargo, las diferencias son más profundas que las planteadas en la definición

del concepto. Justamente en los próximos párrafos se intentará marcar las más destacadas, con el deseo de denotar la importancia de adoptar este nuevo enfoque.

- De la búsqueda empecinada de la venta de un solo producto a la mayor cantidad de personas posibles, se pasa, como lo denota Don Peepers a una actitud concentrada en colocar el mayor número de ellos a un cliente por vez.

- La mezcla comercial compuesta por las cuatro Ps, es sustituida por un concepto más amplio enmarcado dentro de la satisfacción al cliente, la creación del valor y la entrega de un mejor servicio.

- El énfasis en el conocimiento de los mercados genéricos es desplazado por la base de datos que contiene un conocimiento profundo de los clientes.

- Las empresas dejan ya de obsesionarse por diferenciar a sus productos y se comienzan a preocupar seriamente por diferenciar a sus clientes.

- La información comienza a convertirse en el soporte vital e indispensable de la función comercial, en lugar de considerarla como un accesorio de una gestión basada en la Ingeniería de Productos.

- La gestión de ventas deja de ser un objetivo puntual y una relación fugaz para convertirse en un continuo imposible de abandonar.

- Los conceptos que antes eran claves, como la economía de escalas, la obsesión por la participación del mercado y los resultados por líneas, dejan lugar a la economía de segmentos, la retención, la lealtad, el valor y los resultados para el cliente.

- La calidad objetiva basada en patrones internos rígidos, cede posiciones en manos de una calidad percibida en la mente de los clientes, que rige las relaciones del mercado.

- La profunda preocupación por la sensibilidad de parte de los clientes al precio es desplazada por el deseo de establecer relaciones vigorosas que permitan justificar un plus en el mismo.

- De una relación en la que el cliente era el radar que detectaba los productos que producía el fabricante, se pasa a una interacción en la que se escucha al cliente y en sociedad se produce basándose en lo que él mismo requiere.

- La función del marketing se revitaliza, de manera que es toda la organización la que ejercita la gestión y no sólo un departamento logístico dependiente de la Gerencia Comercial, muchas veces aislado de los acontecimientos que ocurren en la Organización. No es sólo uno el que mata la venta. Todos aportan para ejecutarla.

- De un cliente que se ve obligado a buscar los productos se pasa a la gestión en la que se intenta, con cuanto medio disponible existe, llevárselos lo más cerca de su bolsillo.

- La publicidad masiva va dejando cada vez más lugar a los medios de llegada directa al cliente como el telemarketing, los mailing, las visitas concertadas, los e-mails, entre otros.

Esta concepción del marketing le da una amplia probabilidad de éxito sólo a aquellas empresas que establezcan relaciones más profundas y de mayor confianza con los clientes. El factor clave de diferenciación, en un mundo en que los productos mucho se parecen, será el servicio.

Las compañías en países muy desarrollados pierden a más de la mitad de sus clientes cada cinco años, los costos de captura de nuevos clientes crecen en el mismo lapso en forma exponencial, mientras que la venta más fácil de realizar es la que tiene al cliente fiel como eje de la transacción. Argumentos que no hacen más que abonar la necesidad de llevar a la práctica la implementación del marketing relacional.

Customer Relationship Management

Las organizaciones, según se comentara, invierten mucho más dinero en venderle a un nuevo cliente que en colocar sus productos entre los ya disponibles. Es mucho más costoso reconquistar un cliente que ha dejado la empresa que tener satisfechos a los que se poseen desde un primer momento. Es mucho más fácil venderle un producto a uno de los clientes que compone la cartera de los existentes que a uno potencial.

Por otra parte, algunos clientes son más rentables que otros, algunos son poco rentables y por último otros no son rentables, ni nunca lo serán.

La buena noticia es que las técnicas desarrolladas para el procesamiento de datos permiten medir en forma individual los aspectos comentados para cada uno de los clientes y la realidad es que muchas de las organizaciones que se destacan en la lucha híper competitiva, están comenzando a utilizar estas herramientas.

Mediante la migración de modelos de llegada puntual a modelos de llegada continua están intentando crear y mantener una relación de lealtad duradera, a la vez que buscan descubrir la forma más beneficiosa de construir esa relación. El principio en que se basa este nuevo paradigma es que sólo una pequeña porción de los clientes con sentimientos positivos y lealtad, son generados por los productos, y en cambio, muchos son el resultado de la satisfacción de sus deseos plasmada en la manera en que se ofrece el servicio.

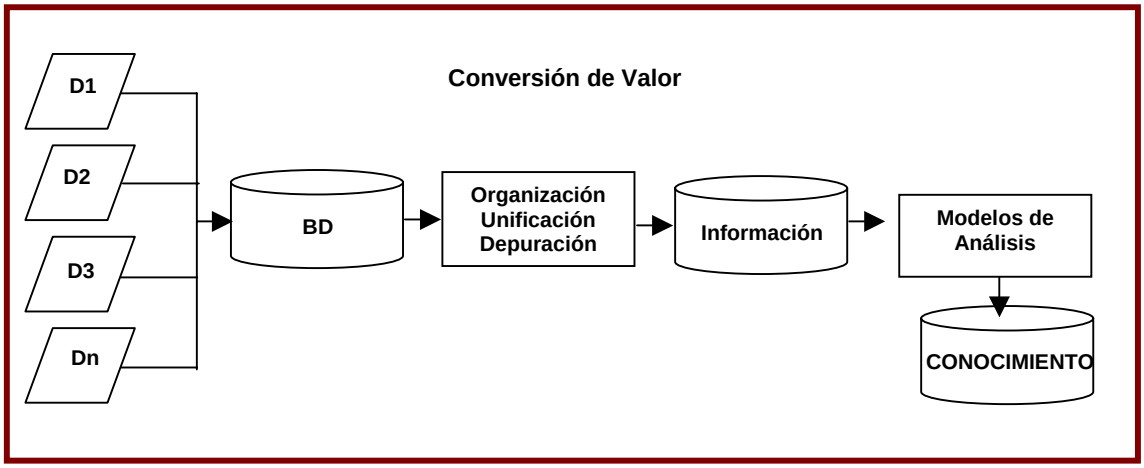
Para que la relación sea lo más beneficiosa posible, las organizaciones comerciales necesitan conocer las preferencias de sus clientes, el estilo de vida de éstos y los hábitos y costumbres que guían sus conductas, de manera de customizar lo máximo posible su oferta.

La metodología que hace posible alcanzar los objetivos de profundo aprovechamiento de la relación con el cliente es denominada **Customer Relationship Management (CRM) y es definida como el proceso de administración de la interacción continua que se registra entre los clientes y la organización.**

Es el conductor de la estrategia dinámica de marketing el principal demandante de este tipo de modelos, dado que tiene que responder en forma continua a preguntas como ¿Cuáles son los clientes más beneficiosos?; ¿Cuál es la razón por la que son los más rentables?; ¿Qué ofertas son las más adecuadas?; ¿Cuál es el perfil de los clientes que son más sensibles a esas ofertas?; ¿Cuáles son los clientes que tienen alta probabilidad de abandonar la empresa?; ¿Dónde están los clientes potenciales con mayor beneficio posible?; para citar, sólo algunas de ellas.

Vale la pena sin embargo, reflexionar acerca de la exactitud de acopio de datos de los clientes de un buen número de empresas, que jamás utilizan esos informes para tomar decisiones. Existe una diferencia importante, en términos económicos, **entre disponer de datos y convertirlo en conocimiento.**

Como lo muestra la **Figura-Conversión de Valor**, debe existir un reservorio de **datos**, capturados a través de diferentes fuentes, que sea organizado, depurado y unificado para convertirlo en **información**, que a su vez mediante modelos de análisis, se transformen en **conocimiento**, **único elemento válido para la toma de decisiones complejas de marketing**.



Teniendo en cuenta esta cadena de conversión de valor de los datos, CRM tiene una arquitectura compuesta por:

- ***Puntos de contacto con el cliente***
- ***Data Warehousing –o almacenamiento de datos-***
- ***Rentabilidad de los clientes***
- ***Data Mining***
- ***Aplicaciones para la relación***

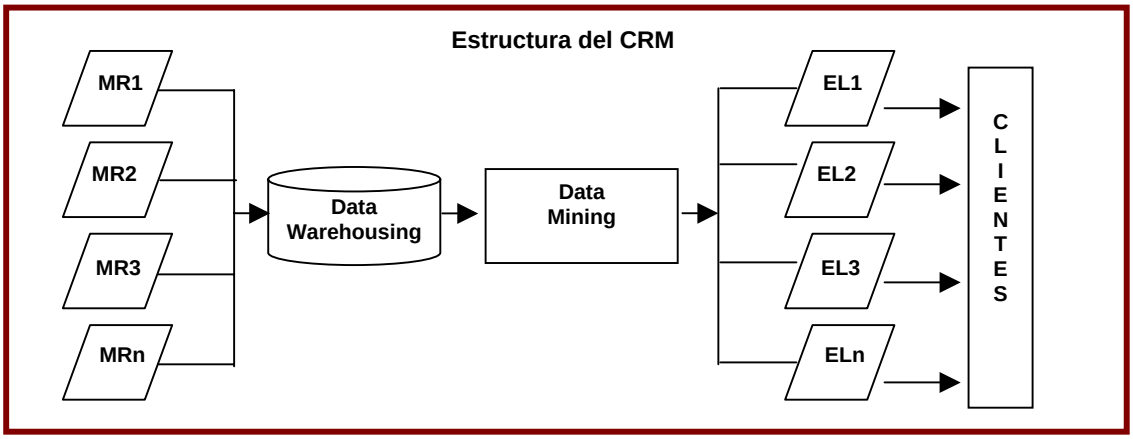
Los cinco elementos interactúan en un todo sistémico y se sinergizan de forma de crear un círculo cerrado de marketing relacional que permite ejecutar tres funciones básicas:

- **Medir**: los resultados de los esfuerzos aplicados al marketing, utilizando como patrón la rentabilidad de los clientes.

- **Predecir**: mediante Data Mining el comportamiento de los clientes y aprender de las experiencias pasadas.

- **Actuar**: utilizando los medios más idóneos para llegar a los clientes con las ofertas más ajustadas a sus deseos.

La minería de datos (Data Mining) se convierte en el componente crítico dentro de CRM y tal vez es el elemento que marca diferencias categóricas con el Marketing de Base de Datos. En esencia se trata de una automatización del proceso de búsqueda y definición de patrones dentro del conjunto de los datos disponibles en la organización (en la mejor situación estos datos se encuentran en el datawarehousing). Se trata de diseñar modelos virtuales de la actividad de los clientes. Estructuras que permiten proyectar su comportamiento. La **Figura-Estructura del CRM**- le da contenido pragmático a la cadena de conversión de datos en conocimiento.



En la figura se simboliza a cada uno de los momentos de contacto con el cliente que permite incorporar datos a la base –identificados como "MR"- y los diferentes medios de comunicación de las estrategias de relacionamiento que producen como consecuencia del análisis del Data Warehouse por parte del Data Mining y que se orientan a f*ortalecer el v*ínculo mediante *la satisfa*cción de los deseos.

Los datos capturados son depurados para luego integrarlos a la base, la cual ya puede contar previamente con entradas similares de otros contactos y momentos de la verdad. La agregación de valor se produce mediante la detección de patrones de comportamiento a través de los indicadores claves de actuación, de manera de poder segmentar, de la forma más adecuada y económica, las acciones a realizar.

Generalmente se identifican tres componentes fundamentales dentro de la estructura de CRM:

- **CRM Operacional**: consistente en la automatización de los procesos en forma horizontalmente integrados, dentro de los que se incluye el contacto personal con el cliente en el momento de la venta, reclamos o consultas, la relación que se establece a través del call center y el contacto mediante los canales de distribución entre otras vinculaciones. Los programas de fidelidad y las mediciones de satisfacción de los clientes son dos herramientas que posibilitan completar con mayor rapidez las bases sobre las que posteriormente trabajará el CRM Analítico.

- **El CRM Analítico**: analiza los datos creados por el CRM operacional con el propósito de mejorar la relación con el cliente. Está indefectiblemente vinculado con la arquitectura del Data Warehouse y con las aplicaciones de Data Marts.

- **CRM Colaborativo**: actúa como un facilitador de la relación entre el cliente y la empresa. Las publicaciones personalizadas, los e-mails enviados a los clientes, la correspondencia emitida por correo, son algunos de los ejemplos de este tipo de Customer Relationship Management.

Los **Indicadores de Actuación** generan un tipo de segmentación básica denominada operativa y varían según el tipo de sector y la actividad en la que la organización se desenvuelve. Sin embargo, existen algunos que son comunes a todas las actividades transaccionales. Entre los Indicadores más habituales se encuentran:

- **Recencia**
- **Frecuencia de compra**
- **Monto comprado por operación**
- **Antigüedad con que opera el cliente**
- **Mezcla de productos comprados**

Una combinación entre los indicadores de actuación y el resto de los datos disponibles convertidos en información, permite obtener cuatro magnitudes de análisis que aportan conocimiento al decididor:

- **COMPORTAMIENTO**: hábitos de compra de los clientes.

- **COMPROMISO**: Análisis de la adhesión de un cliente a una línea o mezcla de productos.

- **ACTIVO**: valor real actual, vitalicio y potencial de cada cliente.

- **DESERCIÓN**: potencial de riesgo, perfil de los abandonadores, debilidades asignadas a la organización e imagen que de ella se tiene.

La mayoría de esta información se convierte en conocimiento a través de la utilización de la multiplicidad de programas disponibles en el mercado, para realizar la minería de datos. Concepto que agrupa a diversos tipos de herramientas, que permiten estudiar la relación entre diferentes variables de un determinado número de casos con disímiles características observadas. En el caso de Data Mining es el propio programa inteligente el que toma la iniciativa y decide, a diferencia de las herramientas relacionales u OLAP que ceden la acción al usuario. El On Line Analytical Processing (OLAP) caracteriza, consecuentemente a un tipo de arquitectura que hace referencia a un sistema de información de decisiones, fundamentado en bases de datos multidimensionales que le posibilitan al usuario manejarse con amplios grados de libertad al decidir. Recientemente, los desarrollos han posibilitado la respuesta en un tiempo real alcanzándose modelos que se denominan On Line Transactionnal Processing (OLTP), que además de proveer una amplia libertad agilizan sustancialmente los procesos de decisión.

Algunas de las herramientas de Data Mining intentan explicar el comportamiento de una variable en función de otras, mientras que otro grupo de técnicas profundizan acerca de la vinculación entre ellas. A los primeros se los conoce como métodos de dependencia y comprenden la regresión múltiple, al análisis de la varianza y covarianza, la medición conjunta, el análisis discriminante y las redes neuronales. Por su lado a los segundos se los llama métodos de interdependencia y agrupan bajo su denominación al análisis factorial, de conglomerados y de estructuras latentes y los árboles de decisión.

Bajo la acepción CRM, en la práctica se agrupan diversidad de aplicaciones relacionadas con la administración de los clientes en un Contact Center, la pantalla con la historia del cliente que se dispara cuando un representante de atención al público recibe un contacto, la encuesta que se realiza para conocer la satisfacción de los clientes o el e-mail que se envía alertando de los nuevos productos lanzados al mercado, para sólo mencionar algunos ejemplos. Si bien estas son actividades que tienen como objetivo mejorar la relación con los clientes, un sistema de CRM es más que una acción en particular.

Es un sistema en que tres elementos, el Data Warehouse, el Data Mining y un Administrador del Sistema, se encuentran relacionados en forma muy estrecha, conformando un núcleo indisoluble. Siendo esto así, si es que se quiere extraer todas las ventajas que provee el marketing de relaciones. El Datawarehousing integra los datos y el CRM los explota utilizándolo de la mejor manera posible en beneficio de la relación con los clientes.

A la hora de construir un CRM deben llevarse a cabo y desarrollarse un conjunto de actividades básicas e inherentes a su adecuado funcionamiento:

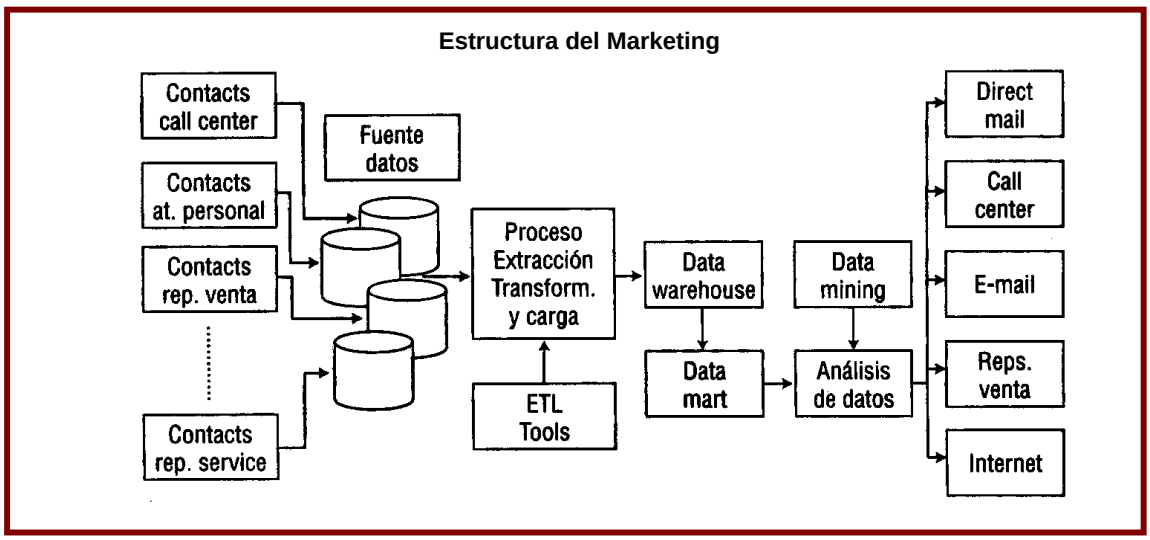
- Diseño de las fuentes de datos.
- Definición de las herramientas de Data Mining a utilizar.
- Generación de una estructura adecuada de aplicación de Data Mining.
- Creación de una plataforma de visualización cómoda y simple para el usuario.
- Administración del proceso en su conjunto.

Seguramente un capítulo se abre siempre que se piense en la integración del CRM con ERP (Enterprise Resource Planning), dado que un nuevo horizonte de expansión y funcionalidad se proyecta a través de toda la empresa. Se trata de una dirección inevitable que sin embargo deberá romper con dos resistencias fundamentales: la primera interna y la segunda externa. En efecto, los principales enemigos endógenos son los grupos individualistas que no quieren cooperar o compartir información, porque suponen que va en detrimento de la rentabilidad de su propio departamento aunque la rentabilidad global se acreciente. La segunda gran fuerza de carácter exógeno es el irresistible deseo de privacidad de los clientes que se resisten frecuentemente a ceder datos que permitan trabajar sobre sus preferencias.

Lo que resulta innegable es que se está en presencia de una estructura de marketing que se funda en un acabado conocimiento de la herramienta informática y de las técnicas de procesamiento de la información.

Nueva Estructura Del Marketing

La nueva estructura del procesamiento comercial se origina en un conjunto variado de contactos que permiten capturar información de los clientes, ya sea mediante llamados al Call center, entrevistas personales realizadas por los ejecutivos de atención, reclamos efectuados por mal funcionamiento de los productos o dudas acerca del funcionamiento. En fin, diversas fuentes de datos alimentan el proceso de alta que, como lo muestra la **Figura-Nueva Estructura del Marketing**, requiere de un proceso de depuración y agrupamiento, para posteriormente ser incorporado a una base de datos común.



Se entiende que el Data Warehouses es un almacén de datos, integrada, con datos no volátiles y específicos del mundo de las decisiones, destinada fundamentalmente a analizar las palancas de negocios potenciales.

Por su lado el Data Mart es una base de datos orientada a un tema específico, puesto a disposición de determinados usuarios en un contexto de decisión descentralizado.

Finalmente, el Data Mining, como ya se comentara, es un conjunto de tecnologías avanzadas que permiten analizar la información de un Data Warehouse de manera de poder explotar tendencias, segmentar o encontrar correlación entre los datos.

El proceso culmina con la generación de nuevas acciones de marketing que utilizan como vehículo de llegada, desde estructuras muy simples, como la atención personal hasta procesos muy complicados como la interacción mediante la Red.

El objetivo es único, cuidar exhaustivamente la relación que se tiene con el cliente de manera de generar ventajas diferenciales como medio de perturbación continuo del equilibrio de un mercado que está en continua mutación, satisfaciendo más y mejor que la competencia.

CAPÍTULO II –
GESTIÓN ESTRATÉGICA DE LA INFORMACIÓN PARA LA BÚSQUEDA DE CONOCIMIENTO

La Gestión Estratégica De La Información

La transición de una economía basada en la producción a una basada en la información ha tenido un profundo efecto no sólo en las operaciones de las organizaciones, sino también en el entorno del análisis de negocios.

Hasta hace poco tiempo, el carácter fundamental de los análisis de datos había permanecido básicamente inmóvil hasta el último gran desplazamiento de mediados de la década del 70, cuando la introducción de los computadores y su evolución a computadores personales provocó un cambio revolucionario en la práctica del análisis de datos. Desde entonces, los analistas de datos han dispuesto de un poder de cálculo suficiente como para eliminar prácticamente todas las restricciones de los diferentes tipos de técnicas estadísticas existentes, y su aplicación en un contexto de negocio ha ganado amplia aceptación y uso. El peso de la educación del analista de datos se ponía en los métodos estadísticos, con énfasis en los fundamentos teóricos y aplicaciones de técnicas. Poco se consideraba la naturaleza o el carácter de los datos, salvo la evaluación de las calidades estadísticas en términos de cumplimiento de los supuestos estadísticos.

Pero la "era de la información" ha traído una segunda revolución, particularmente en el ámbito de los negocios. De nuevo, los analistas de datos se ven desafiados, esta vez reorientados a la aplicación de sus técnicas para un nuevo entorno de investigación y las organizaciones se ven desafiadas por la necesidad de encontrar la información que subyace puertas adentro. Las bases de datos a gran escala con cientos de miles de millones de observaciones ofrecen unos perfiles de atributos detallados durante lapsos de tiempo que pueden durar años. Ahora los datos abundan y las organizaciones buscan nuevas formas de extraer información. Esta segunda revolución en el análisis de datos se basa fundamentalmente en dos tendencias: una avalancha de información y cuestionar la inferencia estadística.

- **La avalancha de la información**

  Las organizaciones de hoy en día se han ido informatizando cada vez más en todas sus áreas funcionales, facilitando no sólo sus operaciones, sino también su habilidad para **recoger datos**. Los avances tecnológicos ofrecen la capacidad de **almacenar efectivamente todos los datos** en formatos comunes y el **acceso** a ellos se encuentra disponible para todos los miembros de la organización. El control de los datos corporativos y formatos de almacenamiento se centraliza en los Data Warehouse o almacenamiento de datos transaccionales ON LINE cuya colección y mantenimiento de integridad se aseguran de manera centralizada. De esta forma los usuarios finales no deben afrontar el problema de integrar múltiples y muchas veces incompatibles bases de datos. Estas tendencias se han combinado para crear algo parecido a una "avalancha de la información", sobrepasando **los métodos manuales de procesamiento e incluso lenguajes convencionales de consulta.** Los warehousing corporativos, por ejemplo, permiten análisis de tipo OLAP (On Line Analytical Procesing), es decir, análisis multidimensionales superiores en complejidad a las consultas tradicionales sobre SQL..

  ¿Cómo puede procesarse esta información y analizarse desde una perspectiva temporal?, ¿Qué técnicas son las más apropiadas para este nuevo desafío?.

  Existe una clara necesidad de disponer de una nueva generación de técnicas con la habilidad de asistir a los usuarios en el análisis de sus datos de manera inteligente y automatizable, que se conoce como Data Mining.
  Estos son algunos de los temas que enfrentan los analistas de datos en este nuevo entorno de investigación.

- **Análisis sin inferencia estadística**

  La avalancha de la información también ha restado atención a las alternativas de la inferencia estadística para evaluar la "significación" de las estimaciones de los parámetros. Pero ¿por qué se querría abandonar este principio, base de la mayoría de las técnicas estadísticas?. En primer lugar, hay un movimiento a favor de "volver a los datos", en el que el analista de datos aplican los menores supuestos posibles para el análisis y "dejan hablar a los datos". El desarrollo de los programas de inteligencia artificial y su progenie (es decir, redes neuronales y algoritmos genéticos) ha llevado a los analistas de datos a explorar un amplio rango de modelos analíticos que no están fundamentados en inferencias estadísticas. Estas técnicas se basan en la aplicación de reglas simples de aprendizaje a un conjunto de datos de la forma menos restringida posible.

Descubrimiento De Conocimientos En Las Bases De Datos (Kdd)

Pocos conceptos han generado tanta atención en la comunidad de la tecnología de la información (IT) como el almacenamiento de datos, y después la extracción de datos. El almacenamiento de datos es el intento de combinar todas las fuentes de información y datos relevantes para una organización en una única base de datos con una estructura susceptible de apoyar el proceso analítico de toma de decisiones a todos los niveles de la organización. La extracción de datos es en cierta forma una reciente perspectiva del análisis de datos con más orientación exploratoria que confirmatoria.

La emergencia del almacenamiento y extracción de datos puede describirse mejor por el dicho "la necesidad es la madre de la invención". Las organizaciones han apoyado de forma entusiasta los aumentos en la automatización e informatización a lo largo de los años y valorado el papel estratégico de la información. Pero se han combinado tres fuerzas para forzar la comprensión de que los nuevos enfoques eran necesarios para la asimilación y uso efectivo de la información disponible en las actuales economías intensivas en información.

El primer factor es la expansión del conjunto de información que casi todas las organizaciones tienen a su disposición. Los sistemas de contabilidad, la automatización de los procesos de fabricación y control de inventarios ofrecen una continua fuente de información. La participación en una economía global y la necesidad de operaciones internacionales exige incluso más cooperación y control, una nueva razón para aumentar la informatización. La aparición de Internet sólo promete un aumento de la oferta y demanda de la información, particularmente a medida que el comercio electrónico ofrece un vínculo directo a los consumidores. La automatización de las fuerzas de venta, fundamentalmente en la informatización y una "oficina virtual", ofrece más de una fuente de información cuya utilización se extiende más allá del vendedor. Las organizaciones han evolucionado desde basarse estrictamente en datos operativos a la integración de información externa dirigida hacia una función específica, sea la interacción de la satisfacción del cliente para el departamento de marketing o los informes financieros para evaluar la evolución del mercado y su rendimiento.

Una segunda fuerza relacionada es que muchas organizaciones, como consecuencia del proceso descrito anteriormente, están produciendo información tan rápido como productos y servicios. Las organizaciones, al capturar los datos del resto de los procesos, están creando un papel paralelo: evaluación constante e inmediata de sus operaciones. Esta fuerza colisionó con la tercera –técnicas que pueden ajustarse a la producción de información con la producción de conocimiento. Lo que se necesitaban eran análisis que se modificaban sin intervención humana, liberando al analista de centrarse en los resultados y el rendimiento del modelo. A medida que estas técnicas se hicieron asequibles y aceptadas, las tres fuerzas se alinearon para promover la acción

La extracción y almacenamiento de datos son elementos complementarios en la mejora del acceso a los datos para la toma de decisiones. El almacenamiento de datos es el mecanismo que facilita el sistema de apoyo a la decisión (DSS), almacenando los datos de la organización

en una única base de datos integrada y ofrece una perspectiva histórica. Los dos conceptos claves que subyacen en el almacenamiento de datos son la integración y la in-varianza en el tiempo. La integración se refiere al diseño unificado de las bases de datos que combinan todas las fuentes de la organización en un único punto de acceso. La in-varianza en el tiempo significa que preserva una perspectiva histórica, de tal forma que tales “trozos de realidad” se encuentren disponibles para cualquier análisis retrospectivo. Una almacenamiento de datos, sin embargo no es una aplicación, sino que facilita las aplicaciones ofreciendo datos desde una perspectiva temporal en el formato adecuado. Su papel como inversión estratégica es cada vez más aceptado.

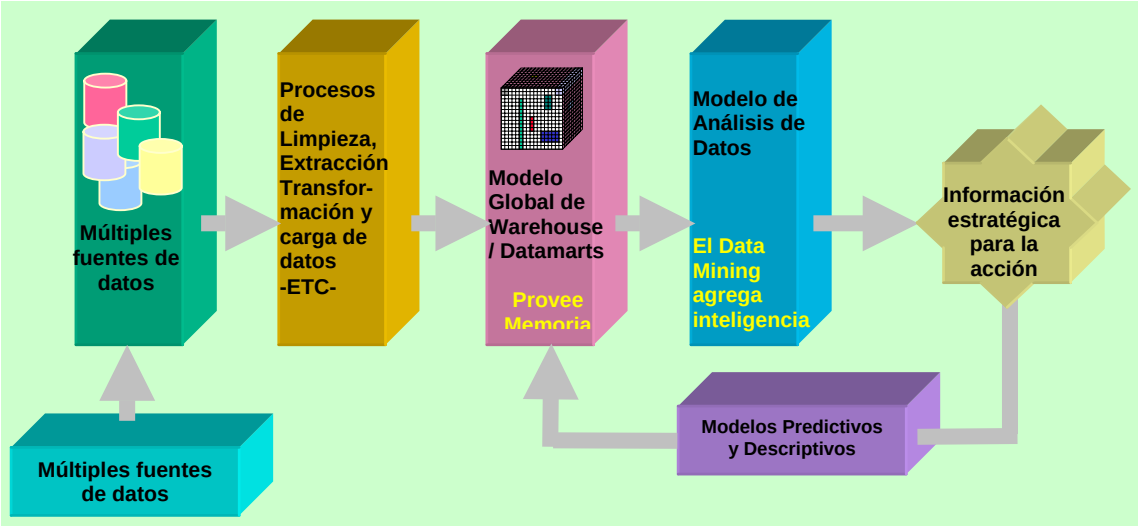
La extracción de datos, también conocida **como descubrimiento de conocimientos en las bases de datos (KDD)** es la búsqueda de relaciones y pautas entre los datos en grandes bases de datos. Como sugiere el término, la extracción de datos tiene una orientación exploratoria de búsqueda de conocimiento oscurecido por las pautas de asociación complejas y la gran cantidad de datos. Sólo se pueden encontrar patrones significativos después de procesar enormes cantidades de datos. Un supuesto implícito es que mediante la revelación de estas relaciones en las bases de datos, los beneficios aumentarán en la medida en que la base de datos refleje perfectamente el entorno de decisión de la organización. Se pueden emplear un gran número de técnicas analíticas en la extracción de datos agrupadas bajo el concepto de Data Mining.

Inteligencia comercial – Business Intelligence-

"Es un paraguas bajo el que se incluye un conjunto de conceptos y metodologías cuya misión consiste en mejorar el proceso de toma de decisiones en los negocios basándose en hechos y sistemas que trabajan con hechos"

Howard Dresner - (Gartner Group),1989

INTELIGENCIA COMERCIAL - RECURSOS Y HERRAMIENTAS



Evolution Business Data to Business Information

Etapa	Pregunta de Negocio	Tecnología disponible	Proveedores	Características
Data Collection (1960)	¿Cuál fue el total de ventas en Capital Federal y GBA?	Computadoras, cintas, discos	IBM, NCR, etc	Retrospectivo Estático
Data Access (1980)	¿Cuáles fueron las ventas por sucursal en Capital Federal y GBA?	RDBMS SQL	Oracle, Informix, Sybase, etc	Retrospectivo Dinámico
Data Navigation (1990)	¿Cuál fue el total de ventas en Capital Federal?Drill down a GBA	OLAP DW	Pilot, Discoverer, Arbor, etc	Retrospectivo Dinámico Niveles múltiples
Data Mining (2000)	¿Cómo evolucionarán las ventas en el próximo año?	Algoritmos avanzados Multiprocesa-dores	Intelligent Miner (IBM), SGI, SPSS, SAS, etc	Prospectivo. Proactivo

Las Tecnologías Para La Búsqueda Del Conocimiento

Introducción Al Data Mining

El "Data Mining" ó Minería de Datos es el proceso estadístico que permite convertir bases de datos en información estratégica. Funciona al explorar grandes cantidades de datos (bases de datos ya existentes) para descubrir patrones, correlaciones y anomalías. El "Data Mining" desarrolla modelos de estos patrones que permiten tanto explicar como predecir comportamientos.

El Data Mining permite responder preguntas que antes no se podían resolver con ningún tipo de análisis, ni siquiera con investigación de mercados.

El término "**Data Mining**" es relativamente reciente y por lo tanto su definición y alcances no están totalmente acotados. En líneas generales, se describe como un **proceso típicamente interactivo e iterativo dedicado a encontrar patrones de comportamiento significativos y no triviales subyacentes en los datos**. Algunos autores incluyen el proceso de Data Mining en otro más general de Búsqueda de Conocimiento o KDD -Knowledge Data Discovery-. El Knowledge Data Discovery es un proceso no trivial de identificación de patrones comprensibles, válidos, nuevos y potencialmente útiles en los datos. El Data Mining es el conjunto de métodos aplicables en el proceso de KDD.- Es importante destacar que estas técnicas de Data Mining deben estar apoyadas por un software inteligente que automatice parte del proceso y oriente al usuario en la obtención de los resultados.

En el año 1997, los autores del libro "Data Mining Techniques for Marketing, Sales and Customer support" [1] enunciaron:

"Data Mining es el proceso de exploración y análisis, por medios automáticos o semiautomáticos, de grandes volúmenes de datos para encontrar patrones significativos y reglas de negocio"

Revisando esta definición años más tarde, concluyeron que: "*nos gusta el énfasis puesto en el tratamiento de grandes volúmenes de datos*"; "*nos gusta la noción de que los patrones deben ser significativos*"; si hay algo que rechazamos es la frase "*por medios automáticos o semiautomáticos*", no porque no sea cierto –sin automatización es imposible minar grandes volúmenes de datos- sino porque entendemos que se ha puesto demasiado énfasis en las técnicas de automatización y no así en las etapas de exploración y de análisis". Esto ha modificado el entendimiento de muchas personas que creen que la minería de datos es un **producto** que puede comprarse, en lugar de una ***disciplina que debe dominarse***.

Data Mining **no es** un fin en sí mismo ni una solución instantánea, sino un proceso que ayuda a encontrar soluciones a los problemas de negocio.

La implementación de modelos teóricos de análisis y minería de datos sacan a la luz oportunidades del negocio que no se ven a simple vista.

La implementación del proceso de Knowledge Data Discovery –y más precisamente modelos de Data Mining- se fundamenta en el hecho de que no se consideran supuestos "a priori" sobre los datos, más bien se deja "hablar a los datos" sin hacer ninguna presunción de los resultados.

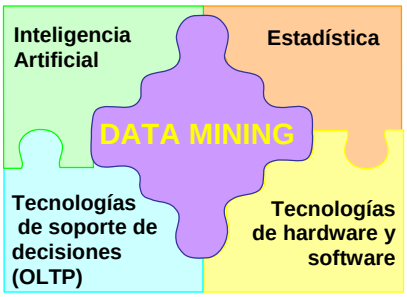
El análisis puede aplicarse en dos sentidos: **dirigido** o **no-dirigido**.

- El **descubrimiento de conocimiento dirigido** intenta explicar o categorizar algún campo de los datos particular como ingreso o respuesta.
- El **descubrimiento de conocimiento no-dirigido** intenta encontrar patrones o similitudes entre los grupos de datos sin el uso de variables designadas ó predefinidas.

[1] Michael Berry, Gordon Linoff ; Data Mining Techniques for Marketing, Sales and Customer Support; Wiley. USA, 1997

Muchas de las herramientas de Data Mining son técnicas tradicionales que previenen de las áreas de la estadística y la inteligencia artificial. Lo que marca la diferencia con los usos convencionales de estas técnicas y el enfoque de Data Mining **es su integración en un contexto de software inteligente que facilita su aplicación y la automatización de parte del proceso interactivo de búsqueda de patrones significativos de comportamiento**.

El Data Mining integra diferentes disciplinas:

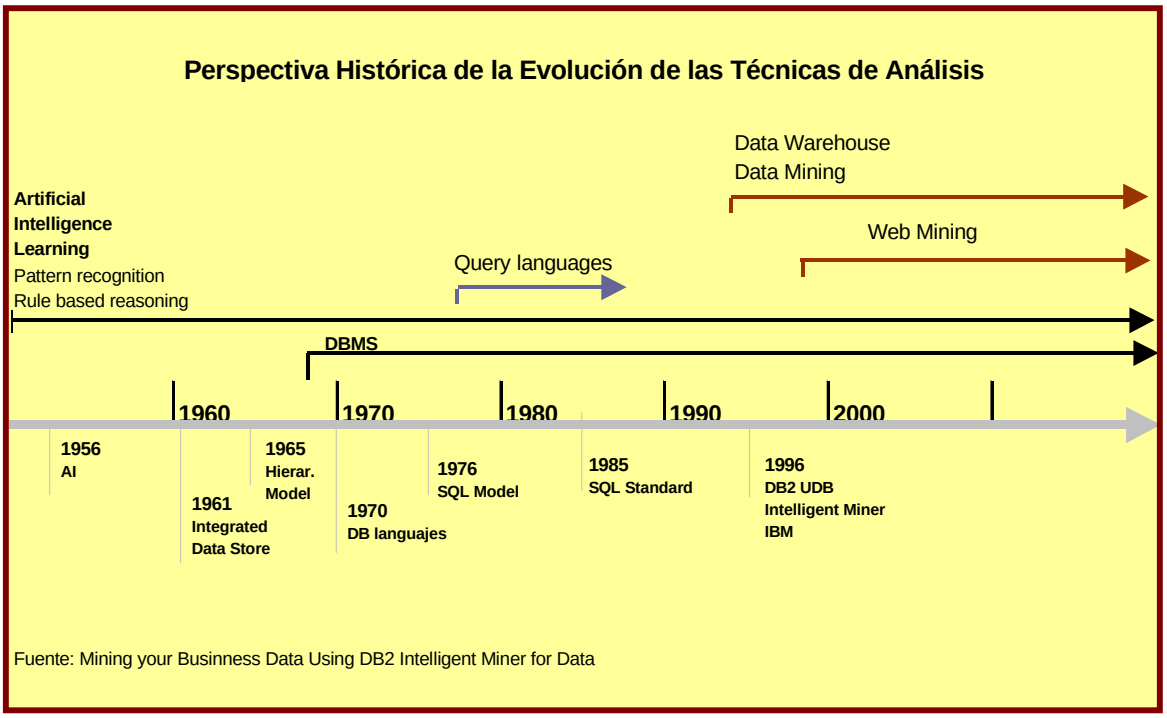


- **Inteligencia Artificial**
  La Inteligencia Artificial se integra a la Minería de Datos a partir de la incorporación de las redes neuronales. Como se verá en el desarrollo del documento, estas redes artificiales se utilizan para construir modelos predictivos no lineales que aprenden a través del entrenamiento y que se asimilan a los modelos de redes de neuronas biológicas.

- **Estadística**
  La estadística es la disciplina que extrae información general a partir de datos específicos. Es el estudio de la estabilidad en la variación. El arte de examinar, sumarizar y extraer conclusiones a partir de los datos.

  Los métodos estadísticos constituyen el "corazón" de muchas técnicas de minería de datos, los cuales originalmente fueron diseñadas con propósitos confirmatorios.

Perspectiva Histórica De La Evolución De Las Técnicas De Análisis



Problemas Típicos A Resolver Con Data Mining

El Data Mining se ha utilizado básicamente en tres áreas:

- **Estrategia y Marketing**
- **Análisis de crédito y fraudes**
- **Mejoramiento de procesos productivos.**

En el área Estrategia y Marketing, el Data Mining utiliza bases de datos de clientes o distribuidores y permite conocer y trabajar con los clientes uno a uno. Permite contestar preguntas críticas como las siguientes:

1. ¿Qué clientes se irán y cómo retenerlos?: análisis de preferencias y vulnerabilidad. Modelos de Predicción de abandono.
2. ¿Qué clientes son más rentables? ¿Cómo concentrar la cartera de clientes en los más rentables?: segmentación de clientes uno a uno con variables como rentabilidad o patrón de consumo. Esta segmentación detallada permite discriminar clientes con promociones, precios y servicios.
3. ¿Cómo hacer más eficiente el gasto promocional? La minería permite seleccionar los prospectos en promociones para aumentar la tasa de respuesta y la efectividad de una campaña. Permite realizar el tracking de campañas y predicción de respuesta / no respuesta.
4. ¿Cuál es el comportamiento de los clientes en la web? ¿Cuál es su perfil de ruteo de páginas y visitas?: Web-Mining, permite desarrollar análisis de tráfico y uso de recursos e-business.
5. Definición de marcos muestrales para investigaciones de mercado y encuestas de customer satisfaction.

La selección de la metodología de Data Mining dependerá del objetivo de problema a resolver, los cuales se pueden agrupar en las categorías que se detallan a continuación. Estos problemas, podrán resolverse con uno u otro algoritmo, en función de la hipótesis a demostrar:

- **Problemas de Clasificación.**
- **Problemas de Clustering o Segmentación.**
- **Problemas de Asociación o Dependencia.**
- **Problemas de Predicción o Regresión.**
- **Problemas de Estimación.**
- **Problemas de Descripción y visualización.**

PROBLEMA DE CLASIFICACIÓN

Es uno de los problemas más comunes a resolver mediante técnicas de Data Mining ya que parece ser un imperativo humano, la necesidad de asignar cada especie en su casta de origen. La clasificación consiste en examinar los rasgos de un objeto dado y reasignarlo dentro de un juego de clases predefinido. El análisis es útil en situaciones donde la muestra total puede dividirse en grupos basándose en una variable dependiente caracterizada por varias clases conocidas. El objetivo primario de la clasificación es entender las diferencias de los grupos y predecir la verosimilitud de que una entidad (persona u objeto) pertenezca a una clase o grupo particular basándose en varias variables métricas independientes. La aplicación del Data Mining en estos caso permite distinguir:

- Innovadores - no innovadores de acuerdo a sus perfiles demográficos y psicográficos.
- Usuarios habituales u ocasionales de un producto, compradores de marcas de ámbito nacional o restringido.
- Clasificación de solicitantes de crédito con bajo, medio o alto riesgo: modelos de scoring.

En todos estos ejemplos, existe un limitado número de clases para asignar cada alternativa dentro de una clase

Las herramientas de minería que típicamente se utilizan para producir estas clasificaciones son:

- Árboles de decisión
- Análisis discriminante

PROBLEMAS DE CLUSTERING O SEGMENTACIÓN

El análisis cluster es una técnica que se utiliza para desarrollar subgrupos significativos de individuos u objetos. De forma específica, el objetivo es clasificar una muestra de entidades (personas u objetos) en un número pequeño de subgrupos mutuamente excluyentes basados en similitudes entre las entidades. En el análisis cluster, a diferencia de los algoritmos de clasificación, los grupos no están predefinidos. Por consiguiente se usa la técnica para identificar los grupos. Bajo un esquema de clasificación los grupos se arman en función a las clases definidas sobre la base del modelo desarrollado como producto de la experiencia y preclasificación de otros ejemplos.

Habitualmente, el análisis cluster implica al menos dos etapas. La primera es la medida de alguna forma de similitud o asociación entre las entidades para determinar cuántos grupos existen en realidad en la muestra. La segunda etapa es describir las personas o variables para determinar su composición.

La segmentación de los datos es la tarea inicial de todo proyecto de Data Mining ya que sirve de soporte a todas las actividades que involucra, por ejemplo, el Ciclo de Relacionamiento con los Clientes.

Desde el punto de vista metodológico esta segmentación puede realizarse a partir de criterios pre-establecidos –segmentos por deciles de facturación- o a partir del análisis conjunto de todas las variables que definen el comportamiento de la unidad de análisis (el cliente, por ejemplo). Este último enfoque se denomina “data driven segmentatios”, es decir, los datos mandan sin recibir supuestos ni restricciones a priori.

En modelos de clustering, las clases no están predefinidas y no hay anteriores al modelo que justifiquen su aplicación. Los segmentos o clusters se conformen de acuerdo a las características similares que los describen. Depende luego del analista, explicar o definir cuál es el perfil y las variables que describen cada segmento. Los atributos de los elementos que componen los clusters dan perfil a cada grupo.

Las herramientas de minería que típicamente se utilizan para producir estos agrupamientos “data driven” son:

- Árboles de decisión
- Algoritmos de clustering

La elección de la herramienta dependerá de los objetivos que se persigan con la segmentación.

Los árboles de decisión se aplican en caso de contar con una variable objetivo, por ejemplo, la tasa de respuesta a un mailing, y cuyo comportamiento se requiere explicar. Esta variable es la que orienta el aprendizaje del modelo y por ese motivo se define a estas técnicas como de “aprendizaje supervisado”. La segmentación intentará separar a los individuos en grupos homogéneos según la variable objetivo.

Por otra parte, los algoritmos de clustering, trabajan de un modo sin supervisión, es decir sin la guía de ninguna variable en particular. Todos los datos entran en el análisis y el éxito consiste en agrupar a los individuos en segmentos que resulten significativos para los objetivos del negocio.

PROBLEMAS DE ASOCIACIÓN O DEPENDENCIA

Los modelos de Asociación constituyen una forma de clustering que trata de encontrar grupos de ítems que ocurren juntos en una transacción y proporciona como resultado las reglas de asociación de los productos que se compran en forma conjunta. Se aplica al Análisis de Canasta de Mercado (Market Basket Analysis).

Los modelos de asociación permiten derivar reglas tales como: cuando un cliente compra el producto A, entonces, también compra el producto B, en el 80% de los casos. Este patrón está presente en el 15% de las transacciones.
Estos análisis permiten identificar oportunidades de cross-selling y lay-out de productos asociados.-

PROBLEMAS DE PREDICCIÓN Ó REGRESIÓN

La regresión múltiple es el método de análisis de apropiado cuando el problema incluye una única variable métrica dependiente que se supone está relacionada con una ó más variables métricas independientes. El objetivo es predecir los cambios en la variable dependiente en respuesta a cambios en varias de las variables independientes. El procedimiento es útil siempre que el interés sea predecir la cantidad o magnitud de la variable dependiente.
Por ejemplo:

- Se pueden predecir los clientes próximos a abandonar la compañía (variable dependiente) con información referente a nivel y frecuencia de consumo y registro de incidentes o reclamos (variables independientes).

- De la misma forma, el investigador puede intentar predecir las ventas de una compañía a partir de información sobre sus gastos de publicidad, el número de vendedores y el número de sucursales que distribuyen sus productos.

La técnica de Análisis de Canasta de Mercado, utilizada para conocer que ítems concurren en forma conjunta en una transacción, también puede adaptarse a un modelo que permita predecir compras o acciones futuras a partir de los datos actuales.

La elección de la técnica depende de la naturaleza del dato de origen, el valor del dato a predecir y la importancia de su aplicación explican la aplicabilidad de la predicción.

PROBLEMAS DE ESTIMACIÓN

A diferencia de la clasificación, que trata de valor discretos (por ejemplo: si/no), la estimación se trata de resultados a través de valores continuos. En la práctica, la estimación es a menudo utilizada para realizar modelos de clasificación. Véase el siguiente ejemplo ilustrativo: Una compañía de tarjetas de crédito está deseando vender espacios de publicidad en sus sobres de facturación a un fabricante de botas de ski podría construir un modelo de clasificación que clasifique a todas sus tarjeta-habientes en una o dos clases: skier ó no-skier. En otra aproximación puede construir un modelo que asigne a cada tarjeta habiente una “propensión para abrir una cuenta para productos de sky”. Este podría ser un valor 0 ó 1 indicando la probabilidad estimada de que el cliente sea skier. La herramienta de clasificación ahora se reduce a establecer un scoring de cuentas. Cualquier cuenta cuyo score sea mayor o igual que el scoring se considera como skier y a cualquiera con menor scoring es considerado como no-skier.

PROBLEMAS DE DESCRIPCIÓN Y VISUALIZACIÓN

Algunas veces el propósito del Data Mining es simplemente describir que está pasando con la información residente en la base de datos, aumentando el conocimiento sobre las personas, productos o los procesos que producen los datos en la primera fase. Una buena descripción de las conductas, a menudo sugerirá una mejor explicación al evento. Por lo menos una buena descripción sugiere dónde comenzar a desarrollar una explicación al evento. Alguno de los procesos descriptos anteriormente tales como Análisis de Canasta de Mercado, por ejemplo, son puramente descriptivos.

El Data Mining también se aplica en los casos en que se requiere distribuir geográficamente los clientes en la zona de influencia de ciertos puntos de venta; para ello se utilizan software de geolocalización (GIS).

Todas estas aplicaciones en una base de datos

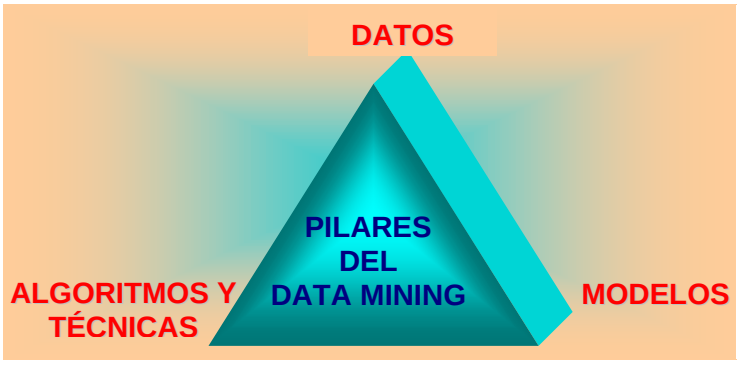
En una aplicación comercial real, el Data Mining debe aplicarse sobre grandes volúmenes de datos, por dos razones:

1. En una base de datos pequeña, es posible encontrar interesantes patrones de conducta y relaciones entre las variables mediante una simple inspección de los resultados a través de simples herramientas como hojas de cálculo y lenguajes de consulta de base de datos multidimensionales, por ejemplo SQL.-

2. La mayoría de las técnicas de Data Mining requieren grandes volúmenes de datos para generar buenas reglas de clasificación, reglas de asociación, cluster, predictores o patrones de comportamiento. Las bases de datos pequeñas llevan a conclusiones inestables basadas en modelos de oportunidad.-

Los tres pilares del Data Mining

Los tres pilares del Data Mining representan las tres áreas centrales necesarias para su implementación:



DATOS

Uno de los pilares del Data Mining son los datos utilizados en el proceso. Sin datos, no hay aplicación posible de ninguna técnica.

El paso inicial para la implementación del proceso de Data Mining es la identificación de las diferentes fuentes y la recopilación de los datos –dentro y fuera de la compañía, si fuera necesario- para resolver los problemas del negocio.

Data Mining utiliza simples vistas de datos, consistentes en tablas con filas y columnas de datos. Algunos algoritmos prefieren no incluir los datos perdidos (missing) o los outliers que son datos fuera de rango que generan un alto impacto en los resultados.

ALGORITMOS Y TÉCNICAS

Las **técnicas** son herramientas utilizadas para extraer información de los datos con el objetivo de resolver los problemas planteados. Existen diversas formas de aplicarlas y cada una de ellas utiliza distintos **algoritmos**.

Los **algoritmos** son secuencias de pasos que describen un camino particular de implementación de la técnica. Los algoritmos son un conjunto de reglas de procedimiento, parecido a una ecuación.

Por ejemplo, “Detección automática de clusters” es la técnica que se implementa utilizando un algoritmo como Simple K-Means, Gaussian K-Means y otros algoritmos.

La aplicación de las técnicas requiere conocimientos matemáticos y estadísticos los cuales, enmarcados en un contexto de negocio, permiten determinar qué técnica se debe aplicar, en qué momento y de qué manera interpretar los resultados.

Las técnicas de Clusterización, Árboles de Decisión y Redes Neuronales son susceptibles de aplicar desde el entorno de una computadora personal, no requiriendo excesiva inversión en hardware para la aplicación de algoritmos de Data Mining.

Por otro lado se requieren analistas y usuarios de negocio que dominen la metodología y sepan cuándo y donde conviene aplicarlo.

MODELO

El tercer pilar del Data Mining consiste en un set de prácticas y habilidades necesarias para construir los modelos predictivos. El foco en este caso son los modelos predictivos –modelos de descubrimiento directo- en lugar modelos de conocimiento indirecto, por dos razones:

Los modelos predictivos se utilizan para encontrar patrones ocultos en los datos. Los modelos indirectos requieren patrones conocidos.

La metodología que construye modelos predictivos está basada en los principios del diseño experimental, ya que necesita conocer todos los factores que afectan al modelo.

Los modelos que se construyen sobre la preclasificación de los datos son conocidos. El proceso de construcción del modelo utilizan los datos para validación y propósitos de testeo.

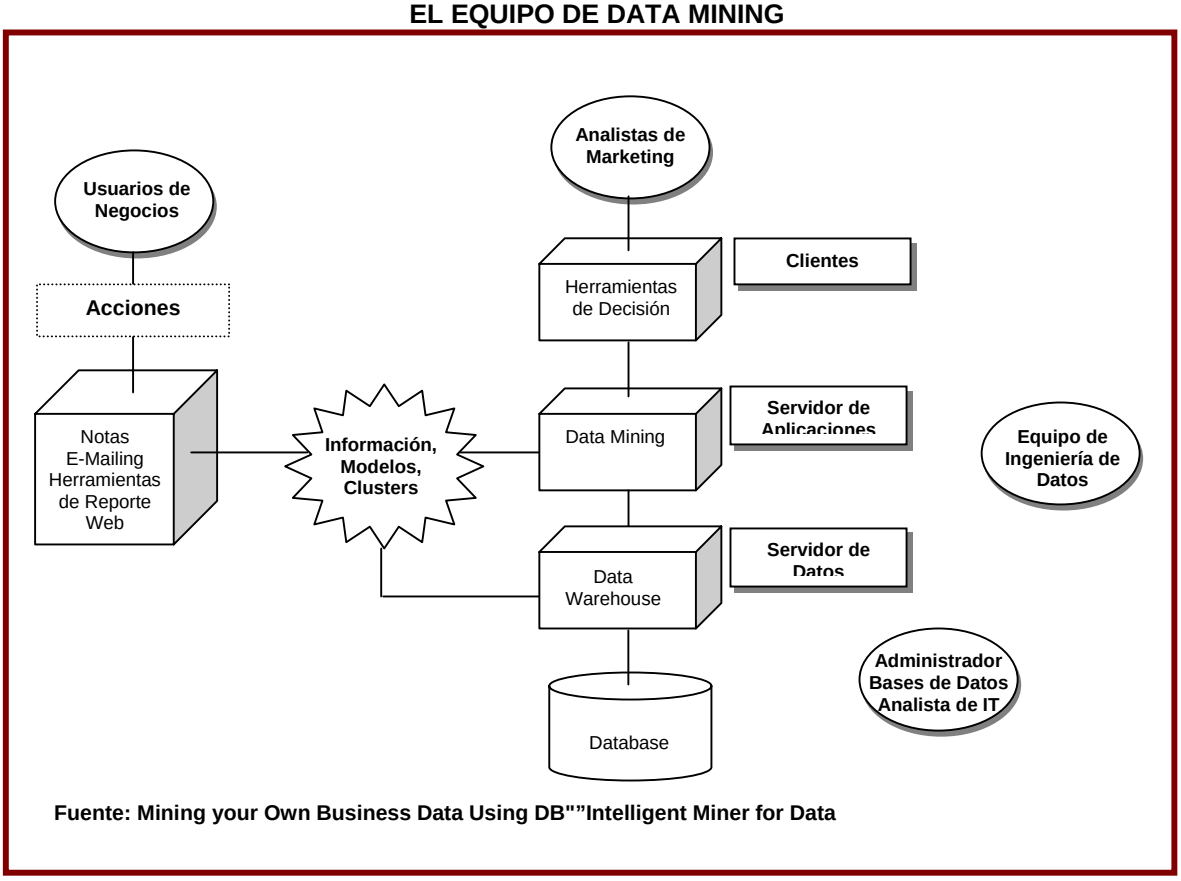
Marco De Ambiente Para Data Mining

Estos tres pilares funcionan en el marco de un **Ambiente de Data Mining:**

El ambiente de Data Mining es la parte o partes de la organización que forman parte del "core" de Data Mining:

- Un **equipo reconocido** por sus habilidades para desarrollar Data Mining.
- **Comunicación entre las distintas unidades de negocio** –sistemas, marketing, áreas comerciales- ya que el trabajo esta enfocado en las necesidades del negocio.
- **Un set de herramientas: hardware y software apropiados.**
- **Acceso de todos los datos necesarios dentro de la organización y habilidades** para publicar los resultados que motivan a la acción.

EL EQUIPO DE DATA MINING



Fuente: Mining your Own Business Data Using DB""Intelligent Miner for Data

El Proceso De Data Mining: Un Círculo Virtuoso

Un proceso de Data Mining consiste básicamente en las siguientes etapas:

EL PROCESO DE DATA MINING

El cumplimiento de los pasos descriptos en el Ciclo del Data Mining constituye la clave para el éxito en la incorporación de la técnica en un proceso de negocio. El resultado de cada una de las etapas es el input para la etapa siguiente y el ciclo completo se desarrolla sobre la base de los resultados obtenidos. Estas etapas se repiten cíclicamente hasta la obtención del modelo final.

1-IDENTIFICAR EL PROBLEMA DE NEGOCIO

El punto de partida para cualquier proceso de Data Mining es definir la hipótesis de investigación (problema u oportunidades de negocio) y los objetivos analíticos en términos conceptuales, antes de especificar cualquier dato o medida. La idea en este punto es ver el problema en términos conceptuales, definiendo los conceptos e identificando las relaciones fundamentales a investigar.

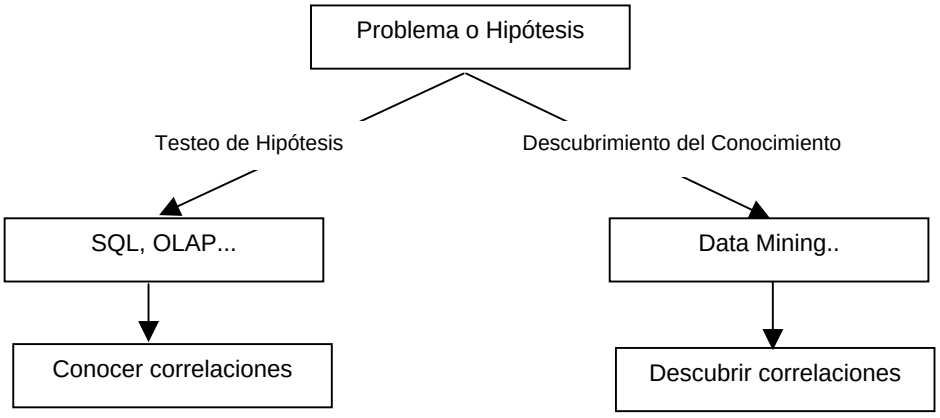
Un modelo conceptual es una simple representación de las relaciones a estudiar. Se definen conceptos, más que variables, se identifican las ideas o temas de interés en primer lugar, sin hacer hincapié en las medidas o modelos a utilizar con posterioridad. Con ello se minimiza la posibilidad de que conceptos relevantes sean omitidos en el esfuerzo de llevar a cabo el proyecto y definir los detalles del diseño.

El desafío es identificar las áreas o situaciones a las que un proceso de Data Mining pueda agregar valor. Existen procesos comerciales que por el escarpado volumen de información que involucran y su relevancia en la organización pueden dar origen a la implementación de un ciclo de Data Mining, por ejemplo:

- Planeamientos de las actividades de marketing para la introducción de nuevos productos.-
- Pricing para productos o servicios
- Segmentación de la cartera / universo de clientes
- Identificación de los clientes en riesgo de pérdida o abandono –attrition / churn-

En estos casos, la oportunidad del negocio es bien entendida o conocida y el Data Mining debe integrarse como parte de ese proceso. En la mayoría de los casos el problema de negocio subyace en las preguntas triviales que los distintos niveles de la organización realizan a menudo y cuyas respuestas están en los datos almacenados en la organización.

Hipótesis de investigación y descubrimiento del conocimiento (Knowledge Discovery)



Existen dos estilos básicos de Data Mining:

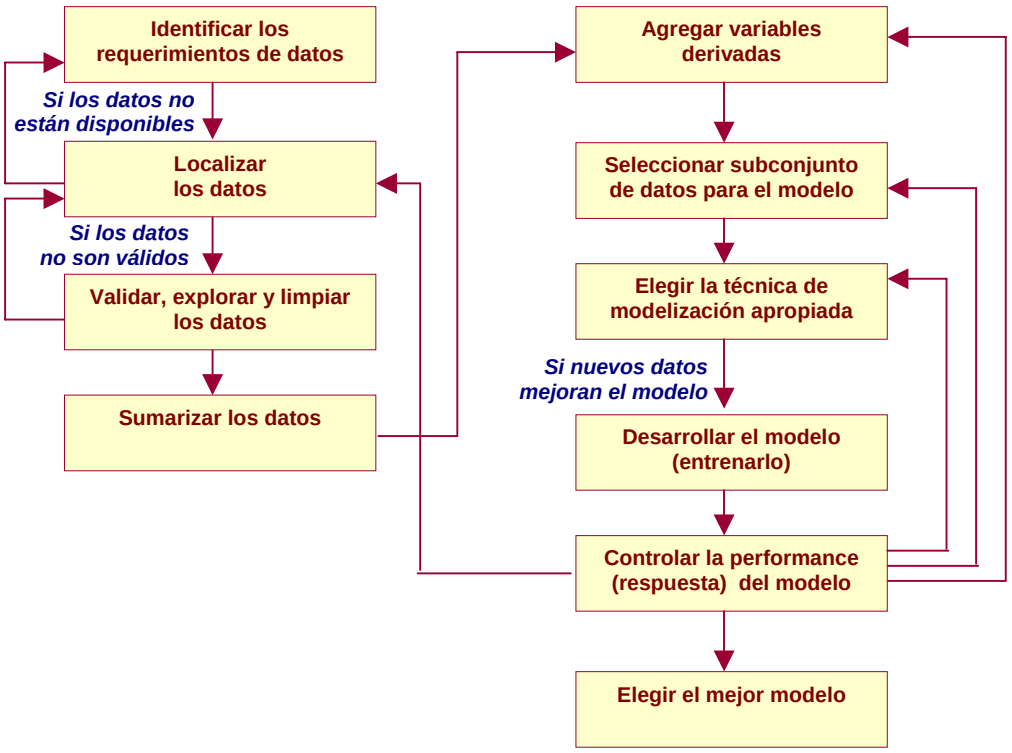
1. **Testeo de Hipótesis**
   Testeo de Hipótesis: es una aproximación “top-down” (de arriba hacia abajo) que intenta probar o refutar ideas preconcebidas. Implica pensar en las posibles explicaciones para la conducta observada y dichas hipótesis dictan los datos a ser analizados.

2. **Descubrimiento del conocimiento**
   Descubrimiento del conocimiento: es una aproximación “bottom-up” (de abajo hacia arriba) que parte de los datos e intenta descubrir o sacar a la luz cuestiones subyacentes en los mismos o que no se tenían por conocidas. Implica dejar a los datos sugerir nuevas hipótesis de testeo.

2-TRANSFORMAR LOS DATOS EN INFORMACIÓN

Es en ésta etapa donde el proceso de Data Mining tiene lugar en sí mismo y donde la información resultante es fundida para producir conocimiento.

TRANSFORMAR DATOS EN INFORMACIÓN



♦ IDENTIFICAR LOS REQUERIMIENTOS DE DATOS

En este punto es necesario evaluar cuáles datos serán necesarios para el testeo de la idea y cuál será el proceso de recopilación de los mismos: consultas a la base de datos, encuestas, información de punto de ventas, información proveniente de sistemas operacionales.

Para cada hipótesis se debe definir la lista de requerimientos de datos, para luego avanzar con la etapa siguiente.

♦ LOCALIZAR LOS DATOS

El proceso de Data Mining requiere de datos. En el mejor de los casos, los datos residirán en el Data Warehouse de la compañía. Sin embargo, a menudo suelen presentarse situaciones donde los datos se encuentran esparcidos en distintos sistemas operacionales, en incompatibles formatos, corriendo en diferentes sistemas operativos y son accedidos a través de herramientas incompatibles entre sí.

La disponibilidad y accesibilidad de las fuentes de datos varían por supuesto, de problema en problema y de industria en industria.

A continuación se detallan algunos ejemplos de las fuentes de datos que pueden utilizarse. El mayor desafío en un proceso de Data Mining es la transformación de los datos en los formatos requeridos por las técnicas y algoritmos de Data Mining.

Los datos provienen de diversas fuentes:

- **Bases de datos Relacionales (RDBMS)**
- **Sistemas operacionales**
  - Sistemas de Puntos de Venta en retail (cupones de sorteos, cupones de descuento, lectura de código de barras)
  - Registros de transacciones de tarjetas de crédito
  - Sistemas de ATMs
  - Sistemas de Depósitos automáticos en bancos
  - Sistemas operacionales de e-commerce
  - Telecomunicaciones - Registros de datos del call center
- Data warehouse
- Data Marts y OLAP
- Otras fuentes:
  - Excel /Archivos ASCII
  - Datos censales / Datos de variables y niveles socioeconómicas macros
  - Formularios de adhesión de clientes: datos demográficos y psicográficos
  - Respuestas de mailings directos
  - Resultados de encuestas
  - Registraciones ad hoc
  - Registros de datos de coberturas médicas

Tan pronto como se tenga acceso a datos provenientes de nuevas fuentes, es recomendable analizar el perfil del mismo a efectos de determinar cuán homogéneo y consistente es. Perfilar el dato incluye realizar sumarizaciones estadísticas campo a campo, contar el número de diferentes valores dados para una categoría de variables y cuando sea apropiado realizar cross tabulaciones entre filas y columnas para determinar la homogeneidad de las variables.

Realizar estos análisis sobre las fuentes de datos, evita incurrir en errores futuros sobre el modelo.

♦ VALIDAR, EXPLORAR Y LIMPIAR LOS DATOS

El proceso de Data Mining integra los datos en el marco de las oportunidades del negocio. La etapa de preparación y limpieza (clean-up) de los datos es la que generalmente insume más tiempo y recursos. En efecto, como los resultados del proceso dependen principalmente de la calidad de los datos, resulta necesario asegurar su validez y consistencia antes de aplicar cualquier algoritmo de modelización.

El análisis de os datos implica la separación, identificación y medida de la variación en un conjunto de variables, tanto entre ellas mismas como entre las variables dependientes y una o más variables independientes. El término clave aquí es "**medida**", dado que no es posible separar o identificar una variación a menos que pueda ser mensurable. La medida es importante para representar con precisión el concepto de interés y es crucial en la selección de la técnica de modelización.

Existen dos tipos básicos de datos:

- **Datos no métricos** (**cualitativos**)

  Los datos no métricos son atributos, características o propiedades categóricas que identifican o describen a un sujeto. Describen diferencias en tipo o clase indicando la presencia o ausencia de una característica o propiedad. Muchas propiedades son discretas porque tienen una característica peculiar que excluye todas las demás características. Por ejemplo, si uno es hombre, no puede ser mujer. No hay cantidad de "género", sólo la condición de ser hombre o ser mujer.

  Las medidas no métricas pueden tener escalas nominales u ordinales. La medida con una escala nominal asigna números que se usan para etiquetar o identificar sujetos u objetos. Las escalas nominales, también conocidas como escalas de categoría proporcionan el número de ocurrencias en cada clase o categoría de la variable objeto de estudio. Por lo tanto, los números o símbolos asignados a los objetos no tienen más significado cuantitativo que indicar la presencia o ausencia del atributo o característica bajo investigación.

  Las escalas ordinales representan un nivel superior de precisión de la medida. Las variables pueden ser ordenadas o clasificadas con escalas ordinales en relación a la cantidad del atributo poseído. Cada subclase puede ser comparada con otra en términos de una relación de "mayor que" o "menor que". Por ejemplo, los diferentes niveles de satisfacción del consumidor individual con diferentes productos nuevos puede ilustrarse en una escala ordinal

- **Datos métricos** (**cuantitativos**)

  Las medidas de datos métricos están constituidas de forma tal que los sujetos pueden ser identificados por diferencias entre grado o cantidad. Las variables medidas métricamente reflejan cantidades relativas o grado. Las medida métricas son las más apropiadas para casos que involucran cantidad o magnitud, tales como el nivel de satisfacción.

  Las escalas de medidas métricas, sean éstas escalas de intervalos y de razón proporcionan el nivel más alto de medida de precisión, permitiendo realizar casi todas las operaciones matemáticas. Estas dos escalas tienen unidades constantes de medidas, de tal forma que las diferencias entre dos puntos adyacentes de cualquier parte de la escala son iguales. La única diferencia real entre las escalas de intervalo y las de razón es que las de intervalo tienen un punto cero arbitrario, mientras que las escalas de razón tienen un punto de cero absoluto.

  Las escalas de razón representan la forma superior de medida de precisión, dado que poseen las ventajas de todas las escalas inferiores más un punto de cero absoluto.

  Es importante entender los diferentes tipos de escalas de medida por dos razones:

  - Para identificar la escala de medida de cada variable empleada de tal forma que no se estén utilizando datos no métricos como si fueran métricos.
  - La escala de medida es crucial para determinar la técnica de modelización apropiada para los datos, consideración hecha tanto para las variables dependientes como para las variables independientes.

  La semántica de los datos debe ayudar para la selección de una conveniente representación y las bondades de la representación elegida gravitan directamente sobre la calidad del modelo y de los resultados posteriores. En algunas oportunidades, la disponibilidad de un datawarehousing bien diseñado minimiza este esfuerzo inicial. En la mayoría de los casos resulta necesario, en cambio, consolidar diferentes fuentes y

formatos de información, eliminar redundancias y duplicaciones, identificar datos faltantes (missings) o fuera de rango (outliers).

Los problemas más frecuentes que suelen presentarse sobre los datos son:

- **Demasiados datos**
  - datos corruptos o con ruido
  - datos redundantes (requieren factorización)
  - datos irrelevantes
  - excesiva cantidad de datos (muestreo)
- **Pocos datos**
  - atributos perdidos (missings)
  - valores perdidos
  - poca cantidad de datos
- **Datos fracturados**
  - datos incompatibles
  - múltiples fuentes de datos

Una vez finalizada la limpieza se deberá seleccionar el conjunto de datos (filas y columnas) sobre los que se aplicará el algoritmo de modelización y pre-procesarlos para, por ejemplo, discretizar valores continuos, ajustar escalas o asignar códigos numéricos a valores no métricos ó valores simbólicos que representan categorías nominales, con el objeto de adecuarlos a la herramienta de Data Mining que se utilizará en el modelo.

Conceptualmente existen cuatro fases distintas en el proceso de preparación de los datos:

1. Un **examen gráfico de la naturaleza de las variables** a analizar y las relaciones que forman las bases del análisis
2. Un proceso de evaluación para entender el impacto que pueden tener los **datos ausentes** sobre el análisis y una serie de alternativas para casos reiterados de ausencia de datos en el análisis
3. Identificación de **casos atípicos**, aquellos casos que por su singularidad pueden distorsionar las relaciones sobre una o más variables analizadas
4. **Métodos analíticos** necesarios para evaluar adecuadamente la capacidad de los datos para cumplir con los objetivos del modelo.
5. **Incorporación de datos no métricos con variables ficticias**

1. Examen gráfico de los datos

Análisis de la distribución de las variables
El punto de partida para entender la naturaleza de cualquier variable es caracterizar la forma de su distribución. Para obtener una perspectiva adecuada de la variable se aplican herramientas como el **histograma**. Con ello se obtiene una representación gráfica de los datos que muestra la frecuencia de los casos (valores de los datos) en categorías de datos.

Análisis de relación entre variables
El método más popular de análisis de las relaciones bivariantes es el **gráfico de dispersión**, un gráfico de puntos de datos basados en dos variables. Se presenta una variable en el eje horizontal y la otra en el eje vertical. Los puntos del gráfico representan los correspondientes valores conjuntos de las variables para cualquier caso dado. El patrón de los puntos representa la relación entre las variables. Cuando los puntos se organizan a lo largo de una línea recta, se tiene una relación lineal de correlación. Un conjunto de puntos curvados puede indicar una relación no lineal. O puede que no existan patrones, sólo un conjunto de puntos aparentemente aleatorios. En ese caso no habría relación.

Análisis de las diferencias entre grupos

Otra de las tareas a la que se enfrenta el analista de datos en éstos procesos es entender el carácter y la diferencia entre dos o más grupos de una variable para dos o más variables métricas. En éstos casos se necesita entender cómo se distribuyen los valores para cada grupo y si existen suficientes diferencias entre ellos como para tener significación estadística. Otro aspecto importante es la detección de casos atípicos que pueden resultar ser aparentes sólo cuando los valores de los datos se separan en grupos. El método que se utiliza para ésta tarea es el **gráfico de cajas** (boxplot), una representación gráfica de la distribución de los datos. Los límites superior e inferior de la caja marcan los cuartiles superior e inferior de la distribución de datos. Por tanto, la longitud de la caja es la distancia entre el primer y el tercer cuartil, de forma que la caja contiene el 50% de los datos centrales de la distribución. La línea dentro de la caja señala la posición de la mediana. Si ésta cae cerca del final de la caja, se indica la presencia de asimetría. Cuanto mayor es la caja, mayor es la extensión de las observaciones. Las líneas que se extienden desde la caja (llamadas bigotes) representan la distancia entre la mayor y la menor de las observaciones que están a menos de un cuartil de la caja. Los casos atípicos son observaciones que se sitúan entre 1.0 y 1.5 cuartiles fuera de la caja.

Los valores extremos son aquellas observaciones mayores que están a 1.5 cuartiles fuera de los límites de la caja.

2. Datos ausentes

Los datos ausentes son habituales en análisis de múltiples variables. El desafío consiste en enfrentarse a los resultados producidos por los datos ausentes en los procesos de estimación y que afectan a la generalidad de los resultados. La ocupación primaria es determinar las razones que subyacen en el dato ausente. Esta necesidad se desprende del hecho de entender el proceso principal de esta ausencia de datos para seleccionar el curso de acción apropiado.

Los efectos de ciertos procesos de ausencia de datos son conocidos e introducidos en el plan de investigación.

Cuando los procesos de ausencia de datos son desconocidos, se intenta identificar cualquier patrón en los datos ausentes que caracterizarían dicho proceso. Al hacerlo se plantean cuestiones como: distribución aleatoria de las observaciones vs. pautas específicas y por otro lado, la relevancia de las observaciones ausentes. Si se encuentran pautas y la extensión de los datos ausentes es suficiente como para garantizar un curso de acción, entonces se asume que está operando algún proceso de ausencia de datos y que alguno de los resultados estadísticos basados en estos datos podrían estar sesgados en la medida en que las variables incluidas en el análisis están influidas por los procesos de pérdida de datos.

El impacto de los datos ausentes es perjudicial no sólo por sus potenciales sesgos "escondidos" sino también por su efecto en el tamaño de la muestra disponible para el análisis. Antes de instrumentar cualquier solución para la ausencia de datos, se debe diagnosticar y comprender los procesos que subyacen en este fenómeno. Algunas veces este proceso se encuentra bajo el poder del investigador y pueden ser identificados explícitamente. En tal sentido, la ausencia de casos se denomina **prescindible**, lo que significa que no se necesitan soluciones específicas para la ausencia de datos dado que los límites de la ausencia de los datos son inherentes a la técnica usada. Un ejemplo de proceso de datos ausentes prescindibles es el "dato ausente" de aquellas observaciones de una población que no están incluidas en la muestra.

La justificación para designar a los datos ausentes como prescindibles es que el proceso de ausencia de datos está operando aleatoriamente (es decir, los valores observados son una muestra aleatoria del conjunto total de valores, observados y perdidos) y que esos efectos aleatorios son identificables y explícitamente ajustados a la técnica usada. No obstante, se debe evaluar la medida y el impacto en que los datos ausentes determinan si es un proceso aleatorio o, en caso contrario, si se puede remediar con alguna de las soluciones existentes.

La ausencia de datos puede ocurrir por muchas razones y en muchas situaciones:

- Factores de procedimiento: errores en la entrada de datos, restricciones de representatividad, fallos al completar formularios.
- Respuestas inaplicables
- Reticencia a brindar datos por parte del encuestado

Aproximaciones al tratamiento de los datos ausentes

Las aproximaciones o soluciones que tratan con los datos ausentes pueden clasificarse en cuatro categorías basadas en la aleatoriedad de los procesos ausentes, en función del método empleado para estimarlos. La aplicación de cualquier otro método introduce sesgos en los resultados.

- Si se encuentran procesos de datos ausentes no aleatorios, se deberá aplicar el método diseñado especialmente al efecto.
- Si se encuentran procesos de datos ausentes aleatorios pueden utilizarse las aproximaciones detalladas a continuación:
  - **Utilizar sólo aquellas observaciones con datos completos**
  - **Supresión de casos o variables**
  - **Métodos de imputación**: proceso de estimación de valores ausentes basado en valores válidos de otras variables y/o casos de la muestra. El objetivo es emplear relaciones conocidas que puedan identificarse en los valores válidos de la muestra para ayudar en la estimación de valores ausentes.

3. Casos atípicos

Los casos atípicos son observaciones de una combinación única de características identificables que les diferencia claramente de las otras observaciones. Los casos atípicos no pueden ser caracterizados categóricamente como benéficos ó problemáticos sino que deben ser contemplados en el contexto del análisis y deben ser evaluados por los tipos de información que pueden proporcionar. Cuando los casos atípicos son benéficos, pueden ser indicativos de las características segmento de la población que se llegaría a descubrir en el curso normal del análisis. Los casos atípicos problemáticos, no son representativos de la población y pueden distorsionar seriamente los tests estadísticos. Debido a la variabilidad en la evaluación de los casos atípicos, se hace imperativo examinar los datos en busca de la presencia de atípicos y la influencia que ejercen.

La ocurrencia de casos atípicos puede clasificarse en cuatro categorías:

- Casos atípicos que surgen de un error de procedimiento (errores en la entrada de datos o codificación).
- Observación que ocurre como consecuencia de un acontecimiento extraordinario. En ese caso existe una explicación para la unicidad de la observación.
- Observaciones extraordinarias para las que no se tiene explicación.
- Observaciones que se sitúan fuera del rango ordinario de valores de cada variable pero que son únicos en su combinación de valores entre las variables.

4. Métodos analíticos necesarios para evaluar adecuadamente la capacidad de los datos para cumplir con los objetivos del modelo.

- **Distribución normal de las variables**
- **Homocedasticidad:** supuesto relativo primordialmente a las relaciones de dependencia entre variables. Se refiere al supuesto de que las variables dependientes exhiban iguales niveles de varianza a lo largo del rango predictor de la variable
- **Linealidad:** supuesto implícito de todas las técnicas multivariantes basadas en medidas de correlación.

5. Datos no métricos como variables ficticias

Un factor crítico en la elección y aplicación de las herramientas de Data Mining es la medición de las propiedades de las variables dependientes e independientes. Algunas técnicas requieren específicamente datos no métricos como variables dependientes o independientes. Sin embargo, en muchos casos, las variables métricas tienen que ser utilizadas como variables independientes. Técnicas como Detección automática de Cluster –Clustering- requieren

variables métricas. El problema se presenta con las variables no métricas: ¿Qué hacer con éstas variables?, ¿se excluyen del análisis las variables como género, estado civil u ocupación?. La respuesta es definitivamente no. Para utilizar estas variables, el analista tiene a disposición un método para usar variables dicotómicas, conocidas como **variables ficticias**, las cuales actúan como variables de sustitución. Una variable ficticia es una variable dicotómica que representa una categoría de variable independiente no métrica. Cualquier variable no métrica puede ser representada como variable ficticia.

♦ **SUMARIZACIÓN DE LOS DATOS, AGREGADO DE VARIABLES DERIVADAS**

La especificación de los datos requeridos para cada técnica en particular será desarrollada en el contexto de la técnica misma. No obstante, a continuación se describirán los problemas más comunes que requieren de la transformación de datos, sin considerar la herramienta utilizada:

- **Sumarización de los datos**
  El nivel apropiado para la sumarización de datos depende del nivel requerido por el análisis: en algunos casos puede requerirse el dato sumarizado y en otros casos puede trabajarse con datos no sumarizados.
  El nivel de sumarización, también está alcanzado por las especificaciones del análisis.

- **Incompatibilidad e inconsistencia de los datos**
  Esta situación es frecuente en datos provenientes de diversas fuentes, donde los mismos datos están expresados de diferentes maneras.

- **Datos de texto**
  En la mayoría de los casos, los datos expresados en modo texto no son útiles para propósitos analíticos: nombre, dirección, etc.
  Cuando la información es relevante, por ejemplo: país, referencia de la fuente, color / modelo, etc. generalmente se trabaja con codificaciones para la variable cuyos valores pueden manejarse en un modelo analítico. Luego, se linquea cada una de las variables del código como las variables para el campo texto.
  Este método se conoce como **variables ficticias**, que actúan como variables de sustitución. La transformación del campo texto dependerá de la herramienta de Data Mining utilizada.

- **Valores faltantes / perdidos ó missing**
  Las fuentes de datos pueden contener valores faltantes comúnmente cuya descripción del contenido usualmente es “nulls”. No todas las herramientas de Data Mining admiten trabajar con éstos valores.

- **Valores fuera de rango ó outliers**
  Las fuentes de datos pueden contener valores fuera de rango y el analista de datos deberá determinar la pertinencia de su inclusión en el análisis.

♦ **SELECCIONAR UN SUBCONJUNTO DE DATOS PARA EL MODELO**

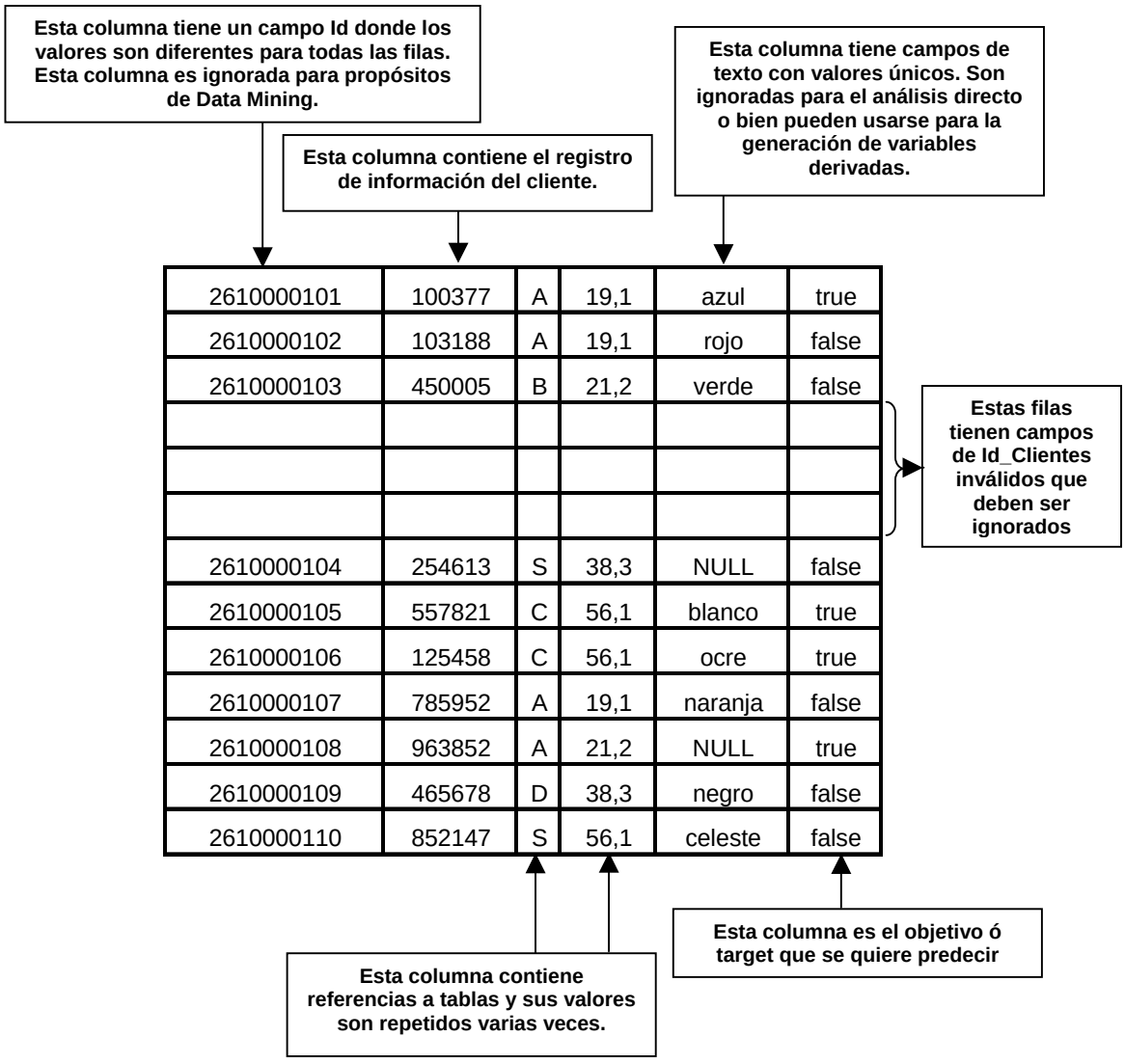
La Forma De Los Datos Para Data Mining

La forma en que se disponen los datos para una aplicación de Data Mining es muy sencilla. Una tabla plana de "n" filas por "n" columnas es necesario y suficiente para la disposición de los datos incluidos en el análisis. El tema clave es qué datos se distribuyen a lo largo de las filas y cuáles datos se distribuyen en las columnas.

Las filas: en las filas se incluye la unidad de acción / análisis o eje de estudio, por ejemplo: un cliente, un ticket. Cada fila corresponde a un caso o a una unidad de análisis. En el caso de clientes, cada fila corresponderá a código de identificación de cliente. Las filas constituyen el nivel de granulidad de los datos.

Las columnas: las columnas representan los datos para cada registro y contiene los atributos de cada unidad de análisis, por ejemplo: frecuencia de uso de una tarjeta de crédito. El rango de la columna refiere a los valores posibles para esas columnas. Los valores de los campos pueden ser numéricos, para lo cual contendrá valores desde los mínimos a los máximos ó categóricos, conteniendo la lista de observaciones posibles dentro de los rangos.

Ejemplo de una Tabla de Datos para Data Mining



Columnas sin variabilidad ó con variabilidad escasa

La distribución más desvirtuada es una columna que posee un solo valor a lo largo de todas las filas. Ese único valor no proporciona información que ayude a distinguir entre diferentes filas. Porque les falta información de contenido, las columnas sin variabilidad deben ser ignoradas para los propósitos del Data Mining.

Los campos de las columnas para los cuales el valor no es asignado o descrito o sólo están designados para valores futuros se completan con valores únicos tales como “null” o “no” o “0”.

Otra causa de columnas con valores unitarios puede presentarse cuando los esfuerzos de Data Mining están enfocados en un subset de clientes. Todos los campos que definen ese subset de clientes pueden contener el mismo valor.

En las columnas que poseen variabilidad escasa, puede apreciarse que casi todos los registros poseen el mismo valor para la columna. En estos casos puede haber algunos casos atípicos o outliers, pero son muy pocos. En estos casos se plantea una disyuntiva respecto si estas columnas deben o no ser definitivamente ignoradas. Para ser ignoradas, los valores deben tener dos características: casi todos los registros deben tener el mismo valor y habiendo unos pocos registros con diferentes valores, éstos constituyan una insignificante porción de los datos. Sin embargo antes de ignorar la columna es recomendable estudiar y conocer el origen y motivo del sesgo y su impacto en el negocio.

Como regla general, si el 95%-99% de los valores de una columna son iguales, es probable que la utilización de la misma no sea útil para el modelo.

Columnas con valores únicos

En el otro extremo, las columnas que tienen un valor distinto para cada fila tampoco suelen ser útiles para el análisis. Ejemplos de columnas con este tipo de información son:

- Nombres de clientes
- Dirección
- Números de Teléfono
- Identificación de cliente
- Identificación del vehículo

Estas columnas no poseen valor de predicción, su contenido únicamente identifica cada registro. Desde el punto de vista de un problema de predicción, estos valores no aportan información para encontrar, mediante la aplicación de herramientas de Data Mining, patrones contenidos en múltiples registros.

Sin embargo, desde el punto de necesidades de visualización o localización geográfica, datos como la dirección y el número de teléfono pueden ser de utilidad para el análisis. La identificación del vehículo puede ser útil para determinar el año, modelo y país de origen. La identificación del cliente secuencialmente asignado, indica cuáles de los clientes son más recientes. Estos son básicamente los casos para los cuales este tipo de columnas puede aportar información: cuestiones de visualización geográfica y recencia de clientes. Fuera de éstas aplicaciones, estas columnas no aportan información de valor y deben ser ignoradas para el modelo.

Columnas que representan sinónimos del target

Cuando la columna es en un todo correlativa con el target, se dice que la columna es sinónimo del target y en ese caso debe ser ignorada. He aquí dos ejemplos:

"Número de cuenta es no-NULL" es sinónimo de los respondentes a la campaña de marketing. Sólo los respondentes que abrieron una cuenta se les ha asignado un número de cuenta.

"La fecha de baja ó perdida es no-NULL" es sinónimo del target si está definido como clientes perdidos o dados de baja.

No obstante, la decisión final de ignorar este tipo de columnas queda en manos del analista de datos, quien debe estar muy al tanto del segmento objetivo de la campaña y el segmento respondente.

En síntesis, las tablas para Data Mining deben considerar las siguientes características:

- **Todos los datos deben mostrarse en una sola tabla o "vista plana" de la base de datos de filas y columnas.**
- **Cada fila debe corresponder a una instancia relevante para el negocio.**
- **Deben ignorarse las columnas que:**
- Poseen **variabilidad escasa a sin variabilidad**.
- Poseen **diferentes valores para cada fila** (en su defecto deben ser reemplazadas por la información contenida en columnas derivadas).
- Poseen **valores únicos para cada registro** de la fila (a menos que se esté tratando de cuestiones de geovisualización).
- Constituyen **sinónimos del target**.

Las columnas cumplen diferentes roles dentro del proceso de Data Mining. Los tres roles fundamentales son:

- **Columnas Input**: usadas como input en el modelo

- **Columnas Target**: usadas sólo cuando se construye modelos predictivos. Contienen información aplicable al modelo como ser propensión a la compra de un producto, probabilidad de responder a una oferta o probabilidad de retener a un cliente. Cuando se construye un modelo descriptivo, no es necesario utilizar esta información.

- **Columnas ignoradas**: columnas no utilizadas en el proceso.

El adecuado nivel de granulidad de los datos

El nivel de granulidad de los datos se refiere a la definición del tamaño de la unidad de análisis para Data Mining. Identificar cuál es la unidad "cliente" en el proceso de análisis. Supóngase el caso de un grupo de tarjetas de crédito en un banco, las posibles definiciones de un cliente pueden ser:

- *Un cliente se corresponde con una cuenta*
- *Un cliente se corresponde con un individuo, independientemente de la cantidad de tarjetas de crédito*
- *Un cliente se corresponde con una familia y contiene todas las tarjetas del crédito sostenidas por todos los individuos en la familia*

Cualquiera de estas apreciaciones es correcta frente a la pregunta de ¿cuál es el cliente?. El propósito aquí no es seleccionar una de ellas, sino ilustrar que existen diferentes niveles de granulidad en el tratamiento de los datos.

- **Cuál es la unidad de tratamiento (fila)**
- **¿Qué es un cliente? Una cuenta, un individuo, una familia**
- **Qué es un producto: un artículo, un rubro, una familia de artículos?**
- **¿Cómo se sumariza la dimensión tiempo? Días, semanas, meses …?**

Extracción Y Almacenamiento De Datos – Una Mirada Sobre El Data Warehouse

El principal objetivo del almacenamiento de datos es la integración de datos de toda la empresa en un formato asequible para el análisis. El almacenamiento de datos persigue al menos tres objetivos:

1. Proporciona ***apoyo para el Sistema Soporte de Decisión, en términos de acceso y organización de los datos.***
2. ***Segrega el acceso al Data Warehouse*** (almacén de datos), reduciendo por tanto la degradación de la evolución de los sistemas operacionales ó transaccionales debido a las peticiones repetidas.
3. ***Fuerza a un reconocimiento de las diferentes estructuras de datos*** necesarias a efectos analíticos frente a los operativos.

Al describir las operaciones del Data Warehouse, es importante distinguir entre dos sistemas: **sistemas operacionales ó transaccionales** y **sistemas de toma de decisiones (DSS)**.

Un **sistema operacional** controla las funciones básicas (transacciones) de los negocios, tales como contabilidad, existencias y gestión de pedidos. Sin estos sistemas, la organización no podría funcionar. Los sistemas operacionales exigen datos que reflejen el status actual de la organización: los datos históricos de escaso interés.

Los **sistemas de apoyo a la decisión** implican solicitudes que tratan con la planificación de la estrategia de la organización. A corto plazo, una organización podría existir sin el DSS. Pero dada la naturaleza crítica de la información, que se constituye como una capacidad de la organización para competir y reaccionar al entorno actual de los negocios, una organización sin el DSS tendría pocas posibilidades de subsistir. Los sistemas operacionales funcionan con sus propias bases de datos sobre la base de la eficacia del procesamiento de múltiples peticiones individuales. Un sistema de apoyo a la decisión, sin embargo, exige la existencia de un Data Warehouse para recoger y convertir los datos a un formato adecuado para las solicitudes DSS. Sin un Data Warehouse, se gasta una enorme cantidad de tiempo y esfuerzo en la preparación de los datos para cada petición de información al DSS.

Data Mining Y Datawarehousing – Una Vista Conectada

Data Warehousing es un conjunto de tecnologías orientado a la efectiva integración de bases de datos operacionales en un entorno que permita el uso estratégico de los datos.

El advenimiento del Data Warehouse permite a las compañías utilizar la información recolectada y obtener un amplio retorno sobre la inversión a la vez que una significativa ventaja competitiva.

De acuerdo con el International Data Corporation (IDC) alguna de las razones que muestran el alto retorno financiero en la implementación del Data Warehouse incluyen los siguientes:

- Capacidad para enfocarse en procesos de negocio y tomar decisiones basados en el conocimiento del sistema completo en lugar de estimaciones con datos incompletos.
- Capacidad de racionalizar y automatizar el proceso de construcción e integración de la información a lo largo de la empresa.
- Costos de hard, soft, almacenamiento en lo referente a desarrollo, implementación y mantenimiento en continuo decrecimiento.
- La capacidad de simultaneidad entre el conocimiento y la administración de los datos, ambos vistos en perspectivas macro y micro dentro de la organización, evitan incurrir en horas de trabajo manual y reducen la ocurrencia de errores resultantes de la asunción de datos incompletos o incorrectos.

♦ ALMACENAMIENTO DE DATOS OPERACIONALES (DATA WAREHOUSE)

Las organizaciones que poseen transacciones on - line - sistemas de procesamiento (sistemas de puntos de venta, por ejemplo) generan datos operacionales. Esos datos son parte de la infraestructura de información de la organización, son detallados, no redundantes y reflejan valores actuales. Con estos datos es posible responder preguntas como "cuánto gastó determinado cliente en un período dado".

La estructura del Data Warehouse por otro lado, es generada sobre la base de un tema o sujeto, un cliente, un producto. El data Warehouse está enfocado a responder a problemas de decisión de la empresa tales como "Cuáles tres productos resultaron pedidos con más frecuencia por la línea telefónica durante un determinado período. Éstos datos generalmente son sumarizados, redundantes dependiendo el punto de vista del dato y más bien estáticos.

En un sistema de operación, un simple registro de dato puede cambiar constantemente considerando los requerimientos del apoyo de decisión, en relación con que el mismo sea guardado en el tiempo como una serie de instantánea de casos de ese registro.

Los datos del Data Warehouse se obtienen de las fuentes de los datos operacionales. Porque el dato operacional es fragmentado e inconsistente, es necesario limpiarlo y darle formatos consistentes, estandarizarlos y proporcionar métodos de acceso en orden a los resultados de los soportes de decisión.

El Data Warehouse tiene por objetivo proveer una arquitectura que proveerá datos corporativos accesible de una manera más rápida y utilizable para analistas y soportes de decisión. Las diferencias de un Data Warehouse con un sistema operacional son las siguientes:

- Organizado por aplicación.
- Soporte periódico / diario del proceso de negocios sobre un nivel de detalle transaccional.
- Constante actualización / up-date.
- Utiliza datos actuales.
- Optimizado para lograr alta performance.
- Acceso limitado por transacción, a menudo acceso directo por clave primaria.
- Soporta alto volumen de transacciones.
- Soporta alto número de usuarios concurrentes.

Aunque la mayoría de las bases de datos de información y operacionales usan la misma tecnología corriendo sobre bases relacionales DBMS, las características de los datos provenientes del Data Warehouse es diferente para los "datos operacionales":

♦ DATOS OPERACIONALES Y EL DATA WAREHOUSE

	Datos Operacionales	Data Warehouse
Contenido	**Valores elementales**	**Datos sumarizados, derivados**
Organización	**Por aplicación**	**Por tema o sujeto**
Estabilidad	**Dinámica**	**Estática hasta su actualización o refresh**
Estructura	**Optimizada para uso transaccional**	**Optimizada para consultas complejas**
Frecuencia de acceso	**Alta**	**Media o baja**
Tipo de Acceso	**Lectura / escritura Actualización campo a campo**	**Lectura / sumarización**
Modalidad de uso	**Predictible / repetitivo**	**A medida / Heurístico**
Tiempo de respuesta	**Segundos**	**Segundos a minutos**

♦ ALMACENAMIENTO DE DATOS OPERACIONALES (OPERATIONAL DATA STORE)

Una interesante variación sobre este tema es una idea de racionalización e integración de los sistemas de operación con el propósito de realizar el soporte de decisión y análisis sobre datos transaccionales. En otras palabras, un Operational Data Store (ODS) es una arquitectura que soporta la decisión de operaciones del día a día y contiene datos corrientes propagados de las aplicaciones de operación. Un ODS provee una alternativa de sistema de soporte de decisión operacional (DDS) aplicaciones que acceden directamente al dato desde los sistemas OLTP, eliminando el impacto de performance que las actividades de DSS pueden tener sobre un sistema de OLTP.

Los sistemas de ODS pueden categorizarse en base a la frecuencia de actualizaciones, como sigue:

- tiempo real o cerca de tiempo real
- periódico
- Diario

Aunque los atributos de ODS son bastante diferentes que un Data Warehouse, las últimas dos categorizaciones hacen a los ODS bastante similar a un Warehouse. Eso es por qué muchos de los requisitos de la aplicación del ODS pueden lograrse directamente a través de acceso bien definido a los datos operacionales o reforzando el proceso de extracción del Data Warehouse.

Sin embargo, algunos desafíos significantes del ODS todavía permanecen. Alguno de ellos son los siguientes:
- Situación de las fuentes de datos apropiadas
- Transformación de las fuentes de datos para satisfacer los requerimientos del modelo de datos ODS
- Complejidad propagación de cambios en tiempo real
- Un sistema de administración de base de datos (DBMS) que combine el procesamiento de consultas –querys- efectivos con el procesamiento transaccional en el marco de las propiedades transaccionales de ACID.
- Un diseño de base de datos optimizado para soportar las más críticas actividades de DSS y, al mismo tiempo, reduciendo el número de índices para minimizar el impacto sobre la performance.

En un típico sistema de Data Warehouse, la información pasa a través de niveles de información, desde fuentes de datos ODS y finalmente almacenamiento en el Data Warehouse. A continuación se verán las diferencias entre bases de datos operacionales, ODS – Operational Data Store- y un Data Warehouse.

	Base de Datos Operacional	**Operational Data Store**	**Data Warehouse**
Contenido	Valores elementales	Valores elementales y nuevos	Datos sumarizados, derivados
Organización	Por aplicación	Por tema o sujeto	Por tema o sujeto
Estabilidad	Dinámica	Dinámica	Estática hasta su actualización o refresh
Estructura	Optimizada para uso transaccional	Optimizada para uso transaccional	Optimizada para consultas complejas
Frecuencia de acceso	Alta	Alta a Media	Media o baja
Actualización	Actualizado campo a campo	Ninguna actualización	Accedido y manipulado: sin actualización directa
Acceso al Dato	Varios registros por transacción	Varios registros por transacción	Algunos registros por transacción

	Base de Datos Operacional	Operational Data Store	Data Warehouse
Tipo de Acceso	Lectura / escritura Actualización campo a campo	Ninguna actualización	Lectura / sumarización
Modalidad de uso	Predictible / repetitivo	Predictible / repetitivo / A medida	Proceso analítico: DSS con rango ancho de datos para discernir tendencias
Tiempo de respuesta	Segundos	Segundos	Segundos a minutos y a veces horas
Performance requerida	Alta	Moderada a alta	Moderada

- **DATA WAREHOUSE: DEFINICIÓN Y CARACTERÍSTICAS**

Un Data Warehouse puede verse como un sistema de información con los siguientes atributos:

- Es una base de datos designada para tareas analíticas, usando los datos para múltiples aplicaciones.
- Generalmente soporta reducido número de usuarios con interacciones largas.
- Su uso es de lectura intensiva.
- El contenido es periódicamente actualizado (principalmente las sumas).
- El contenido corriente y los datos históricos proveen una perspectiva histórica de la información.
- Cada consulta –query- produce resultados largos y frecuentemente involucra varia tablas y múltiples vinculaciones –joins-.

Una definición formal de Data Warehouse es ofrecida por W.H. Inmon:

"Un almacenamiento orientado a un sujeto, integrado, variable en el tiempo, colección de datos no-volátiles en apoyo de decisiones de dirección"

En otras palabras, un Data Warehouse combina lo siguiente:

- Una o más herramientas para extraer campos provenientes de cualquier tipo de estructura de datos, incluyendo datos externos.
- Sintetiza los datos a través de los conceptos de integración, volatilidad, orientado al sujeto.

Los términos frecuentemente asociados a Data Warehouse son:

"Current detail data" – (detalle de datos actuales)
El dato es adquirido directamente de las bases de datos operacionales y a menudo concentra la integridad de la empresa (toda la información de la empresa se concentra el Data Warehouse).

"Old detail data" – (detalle de datos históricos)
Concentra el detalle de datos históricos reales de las áreas involucradas, lo que permite el análisis de tendencias.

"Data Mart"
Técnicamente es una implementación del Data Warehouse, con un alcance más limitado en relación con el Warehouse de la empresa, orientado a una finalidad específica del negocio: marketing, finanzas, producción. Contiene sumarizaciones de datos por departamento y permite customizar los datos satisfacer necesidades particulares de un departamento.

El término se utiliza también para identificar soluciones alternativas a un DW corporativo más reducidas y de menor costo y tiempo de implantación En el proceso de implementación del Data Warehouse, a lo largo de una gran empresa, el Data Mart es el camino para la construcción definitiva del Warehouse de un modo secuencial y de aproximación por fases. Una colección de Data Marts componen el Warehouse de la empresa a lo largo y ancho de la misma.

Summarized data – (datos sumarizados)
Procesos que permiten incorporar datos nuevos a los históricos y actuales para la construcción de reportes y análisis de tendencias.

Drill – down
Es una metodología que permite atravesar las sumarizaciones de datos mediante un análisis top-down, desde la información general hacia la información particular. Por ejemplo: en una sumarización de datos de ventas por zonas geográficas se indica una reducción en los volúmenes de un área en particular; el "drill down" permite al analista analizar el estado, país, ciudad con los registros más desfavorables.

Metadata
Metadata es la información acerca de los datos: provee al usuario la información para facilitarles el acceso e interpretación de los contenidos del data Warehouse: localización y descripción de los componentes del sistema; nombres; definición; estructura; contenidos del Data Warehouse y usuarios finales; identificación de las fuentes de datos; integración y transformación de las reglas del Data Warehouse, registro de actualizaciones históricas; métricas utilizadas para analizar la performance del Warehouse; seguridad de accesos y más.

Sin bien los Data Warehouse son relativos a las diferentes necesidades de las compañías, existen similitudes que permiten clasificarlos y describirlos:

- El Data Warehouse provee un mecanismo de separación de la información operacional – transaccional- respecto de la información procesado que reside en él. Porque el Warehouse es poblado por datos creados por el ambiente operacional, el flujo de información es usualmente un camino para la operación de almacenamiento en el Warehouse.

- Esta es una perspectiva holística que elimina la visión vertical y provee una perspectiva de negocios a lo largo de toda la empresa. El Data Warehouse está previsto para resolver las inconsistencias en el formato, semántica y uso de los datos a través de los múltiples sistemas de operación.

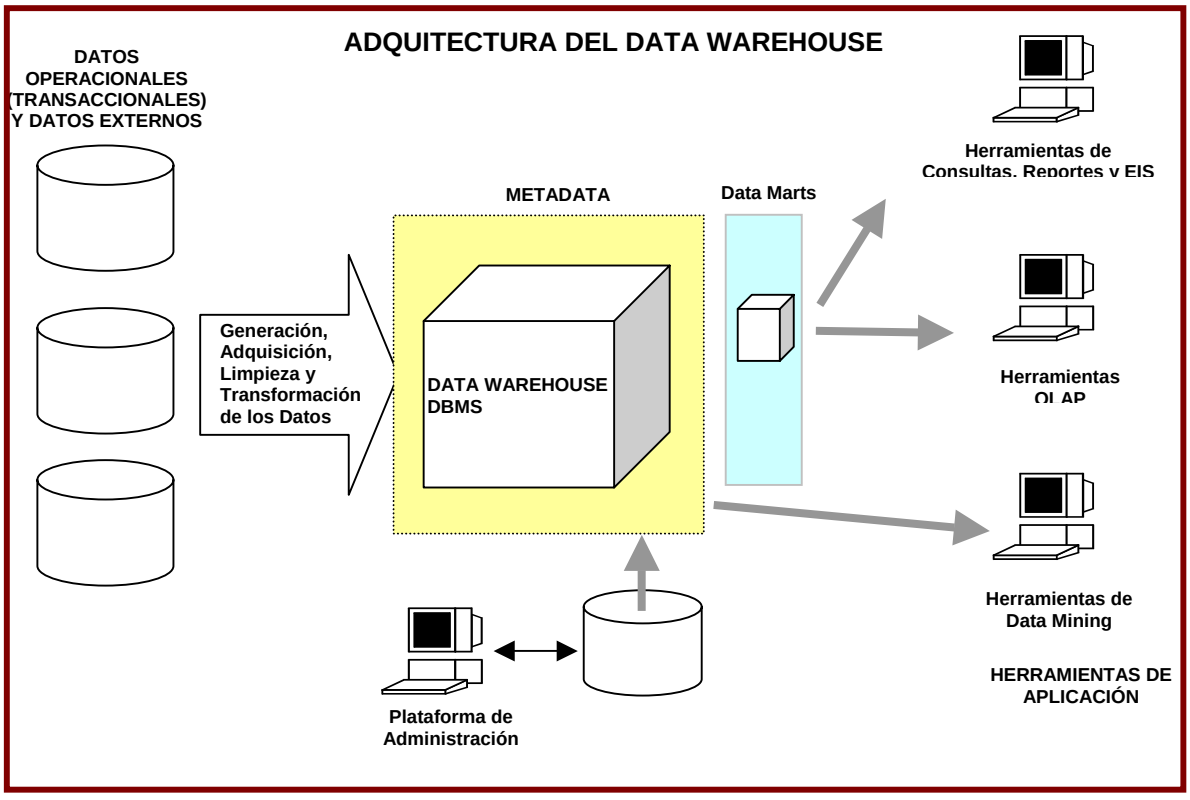
- Parte de las funciones del Warehouse incluyen el procesamiento de los datos desde el formato crudo en las bases de datos operacionales. Los procedimientos del Data Warehouse incluyen agregados, sumarizaciones de los datos haciéndolo más relevante y útil para los usuarios.

- El dato contenido en el Warehouse es un subconjunto de todos los datos contenidos en la organización. El Data Warehouse en considerado como el conjunto universal de los datos provenientes de la operación dentro y fuera de la empresa.

- Frecuentemente los datos externos de la compañía contribuyen a los procesos de decisión. Incorporar los datos externos y mapearlos en las aplicaciones apropiadas es una importante función de Warehouse, transparente para el usuario.

♦ ARQUITECTURA DEL DATA WAREHOUSE

La arquitectura del data Warehouse está basada en sistemas de bases de datos relacionales, es un sistema que funciona como repositorio central para los datos. En la arquitectura del Data Warehouse, el dato operacional y el procesamiento son completamente separados del procesamiento del Data Warehouse.



La fuente de datos del Warehouse son las aplicaciones operacionales. Como el dato ingresa en el Data Warehouse, el mismo es transformado en una estructura y formato integrado. El proceso de transformación implica conversión; sumarización; filtrado y condensación de los datos. Como el Data Warehouse es un almacenamiento que contiene componentes históricos a lo largo del tiempo, debe tener la capacidad de administrar largos volúmenes de datos, tanto como diferentes estructuras de datos y al mismo tiempo.

En el contexto de aplicaciones de Data Mining, ODS –Operational Data Store- es considerado un importante componente en el entorno global del data warehousing, especialmente en la relación entre el Data Mining y similares actividades de análisis. El concepto de la arquitectura ODS/Data Warehouse se muestra a continuación:

En relación con el Data Warehouse, ODS puede ser usado como un área de datos inicial para las fuentes de datos del Warehouse. Recíprocamente, el ODS no tiene que actuar como los datos que organizan el área de ingreso de datos para el Warehouse, sobre todo si éste necesita adquirir datos de fuentes externas que no pueden encontrarse en el ODS. En este caso, ODS y fuentes externas pueden alimentar el Warehouse.

- **OPERACIÓN DEL DATA WAREHOUSE**

Data Warehouse Database

La base de datos central es la piedra angular en el entorno de datawarehousing. Esta base de datos casi siempre es implementada sobre un esquema de base de datos relacional (RDBMS- Relational Database Managment System). Sin embargo, la implementación de un Warehouse basada en tecnología de RDBMS tradicional está a menudo **forzada** por el hecho que se perfeccionan aplicaciones de RDBMS tradicionales para el proceso de bases de datos transaccionales. Ciertamente los atributos del Data Warehouse como el volumen de datos, el procesamiento de consultas ad hoc y las necesidades de generación de distintas vistas para el usuario incluyeron agregados: tablas de múltiples vinculaciones, drill-downs, que condujeron a la convergencia de diferentes tecnologías en el data Warehouse.

Generación, Adquisición, Limpieza Y Transformación De Los Datos

Una significativa porción del esfuerzo de implementación del Data Warehouse se produce en la extracción de datos desde los sistemas operacionales y ponerlos en el formato conveniente para las aplicaciones que correrán bajo el Data Warehouse.
Las herramientas de generación, limpieza, transformación y migración de datos realizan todas las conversiones, sumarizaciones, cambios de estructura y condensaciones necesarias para transformar el dato en información que pueda ser utilizada como herramientas soporte de decisión.

Principales funcionalidades:

- Remover datos no necesarios de las fuentes operacionales
- Consolidar representaciones de datos de diferentes fuentes
- Calcular sumarizaciones y variables derivadas
- Resolver problemas de missings y outliers

Los problemas más frecuentes que suelen presentarse son:

- *Heterogeneidad de bases de datos*: modelos, lenguajes, operaciones, integridad
- *Heterogeneidad de los datos*: diferentes atributos, definición, modalidad de uso.

Metadata

Como se ha dicho anteriormente, este componente provee a los usuarios de información para facilitarles el acceso e interpretación del contenido del DW. Es utilizado para la construcción, mantenimiento, administración y uso del Data Warehouse.

Técnicamente contiene la siguiente información:

- Fuentes de datos.
- Descripción de operaciones de transformación: métodos y algoritmos utilizados para convertir / transformar datos.
- Objeto del Warehouse y definición de la estructura de datos.
- Reglas utilizadas para la limpieza de los datos –data clean up-.
- Mapeo de los datos desde que son capturados del sistema fuente de datos e ingresados al Warehouse.
- Referencias históricas: acceso, back-up, archivo, información entregada, adquisición de los datos, acceso a los datos.

♦ **HERRAMIENTAS DE EXPLOTACIÓN DEL DATA WAREHOUSE**

Herramientas De Visualización

El extenso volumen de datos contenido en un Data Warehouse es imposible de manejar y/o absorber por el usuario final, sin la aplicación de herramientas que permitan visualizar la información en dimensiones asequibles.

Estas herramientas proporcionan un "mapeo" entre los espacios multidimensionales de los datos - o de los resultados de un modelo - y el espacio bi-dimensional de la pantalla de la computadora.

Las herramientas de visualización son utilizadas frecuentemente en aplicaciones de Data Mining:

❖ En la primera fase, visualizar la "montaña de datos" provee al usuario alguna idea de por dónde empezar a minar.

❖ Sobre la marcha permiten mostrar los resultados del Data Mining y los modelos predictivos de una manera entendible para el usuario final.

❖ Su aplicación es efectiva para la construcción de modelos descriptivos y retrospectivos para confirmar o rechazar hipótesis previas del usuario

Las herramientas de visualización pueden dividirse en los siguientes grupos:

a. Herramientas De Consulta Y Reporte –Query And Reporting-
b. Herramientas De Desarrollo De Aplicaciones
c. Herramientas De Procesamiento Analítico On Line (Olap)
d. Herramientas De Data Mining

a. Herramientas De Consulta Y Reporte –Query And Reporting-

Se dividen en dos categorías:

- **Herramientas de reporting**

- Herramientas para la generación automática o asistida de informes periódicos de indicadores de negocio
- Se complementan generalmente con una interfase de asistencia al usuario no técnico para la formulación de querys

- **Herramientas de consulta**

- Estas herramientas "escudan" a los usuarios finales de las complejidades de SQL y las estructuras de la base insertando una interfase de operación entre los usuarios y la base. Permite que los usuarios finales puedan realizar consultas "simples" a la base de datos (identificación de segmentos, listas de mailing, consultas por zonas geográficas,

etc.), generar reportes de una manera mucha más ágil y sencilla. Esta herramienta pierde eficiencia a medida que aumenta la complejidad de las consultas a la base.

b. Herramientas De Desarrollo De Aplicaciones

Las necesidades de análisis de la comunidad de usuarios del data Warehouse a menudo puede exceder las capacidades de creación de las herramientas de Consulta y reportes. En estos casos, se requiere de un complejo set de querys y modelos de datos que el usuario de negocios no tiene a su alcance necesita volver al SQL y/o expertos modelos de datos.

Para esos casos, organizaciones dedicadas al desarrollo específico de herramientas de bases de datos han desarrollado aplicaciones que se integran sobre plataformas OLAP y permiten acceder mejor al sistema de base de datos, incluyendo Oracle, Sybase, Informix. A título de ejemplo cabe mencionar herramientas como: PowerBuilder de PowerSoft, Visual Basic de Microsoft.

c. Herramientas De Procesamiento Analítico On Line (OLAP)

Herramientas basadas en el concepto de bases de datos multidimensionales.
Permiten la elaboración de vistas multidimensionales del DW para optimizar performance
Están soportadas por motores de administración del DW que admiten la construcción de estos “cubos”.

Aplicaciones de negocio típicas por estas herramientas incluyen: estudios de situación y rentabilidad, efectividad en los programas de venta o campañas de marketing, planeamiento de ventas, planificación de capacidad.

Las herramientas OLAP pueden clasificarse en:

- Herramientas Multidimensionales o MOLAP - operan sobre la base multidimensional-
- Herramientas Dimensionales o ROLAP - que operan directamente sobre la base relacional-
- Herramientas Hybrid o HOLAP –que combina habilidades de los grupos anteriores-

d. Herramientas De Data Mining

Las herramientas de Data Mining permiten encontrar patrones no evidentes en los grandes volúmenes de información del DW y proponer modelos predictivos. Las herramientas de Data Mining llegan más allá de las capacidades de las herramientas OLAP, especialmente por ésta posibilidad de construir modelos predictivos en lugar de modelos retrospectivos.

Arquitectura Cliente –Servidor

Para la perspectiva del usuario final, el Data Warehouse representa una fuente de datos, que posibilita la generación de querys o consultas, reportes, análisis y herramientas de análisis de datos como el Data Mining. En otras palabras, el Data Warehouse puede verse como un servidor a los usuarios finales.

Específicamente, las funciones de cliente incluyen interfases de usuario, especificación de querys, análisis de datos, formación de reportes, agregación y acceso a datos. El servidor de data Warehouse desarrolla lógica de datos, servicios de datos, metadatos y almacenamiento.

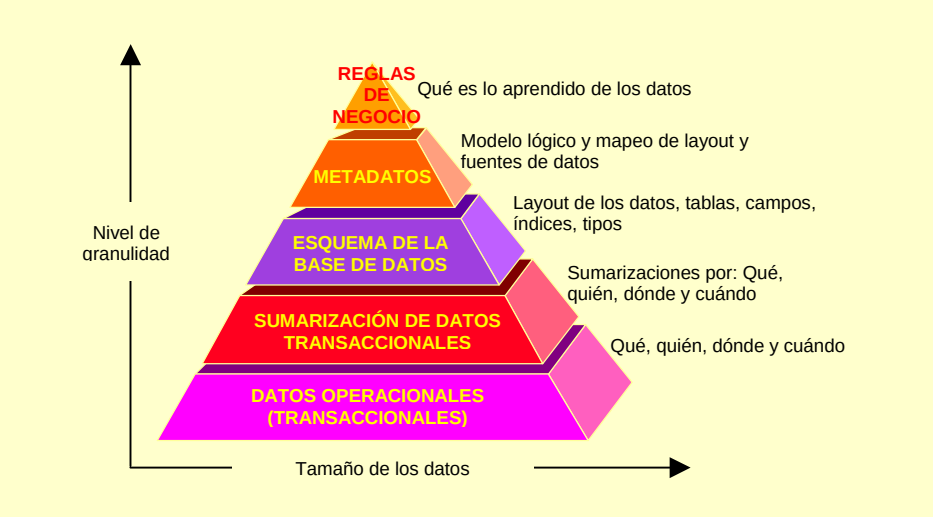
♦ **ARQUITECTURA DE LOS DATOS**

Los beneficios de un Data Warehouse no sólo provienen de sus operaciones día a día y del almacenamiento de datos, sino también de la definición de los datos. Tanto si se definen sus características básicas o a nivel de agregación, como si se forma un perfil completo, estas definiciones de datos son instrumentales en la eficacia y efectividad del Data Warehouse. Cabe aclarar que alguna de éstas definiciones han sido ya suministradas en los puntos anteriores, pero a esta altura se intenta dar una visión estructural desde el punto de vista de los datos.

Las definiciones más importantes son:

- **Datos operacionales frente a datos analíticos**
- **Datos originales frente a datos agregados (sumarización de datos)**
- **Esquema de la base de datos**
- **Metadatos**
- **Reglas de negocio**

LA ARQUITECTURA DE LOS DATOS



- **Datos operacionales frente a datos analíticos**

  Los datos operacionales o transaccionales constituyen el fundamento de mantenimiento automatizado de las operaciones día a día en una organización. Este tipo de datos constituye la forma más elemental de los datos: transacciones de compra, transacciones de tarjetas de crédito. Estos datos no se ajustan bien a un análisis profundo. Un interés central del almacenamiento de los datos es la creación de los datos analíticos mediante el procesamiento e integración de datos operativos en la base de datos en un formato adecuado a las solicitudes de extracción de datos. Los datos analíticos reflejan una estructura interna que combina eventos separados alrededor de un objeto común, tanto si se trata de la historia de compras de un cliente como las ventas de un producto u otra unidad de análisis. Estos procesos de transformación ponen de manifiesto que el almacenamiento de datos no sólo consiste en el depósito de datos, sino también en como se almacenan.

- **Datos originales frente a datos agregados**

  Una segunda característica de un almacenamiento de datos es su capacidad para resumir **datos originales** y almacenarlos como **datos agregados** a un nivel superior de análisis. Los sistemas operacionales descansan en los datos originales dado que se centran sólo en eventos únicos, no en compuesto de eventos. Las aplicaciones de DSS, sin embargo, pueden beneficiarse del aumento de la velocidad de acceso en la oferta de datos agregados. Este punto es crucial a medida que la base de datos aumenta de tamaño, haciendo que las peticiones que requieren resúmenes de datos sean menos eficientes. Además, la rápida respuesta de OLAP se desprende de su acceso a los datos agregados, evitando los retardos asociados con el proceso de resumirlos. La clave está en agregar la información que se necesita, dado que los datos agregados pueden incorporarse en gran medida a los requisitos del almacenamiento. A medida que la dimensión de la base de datos y la complejidad aumentan, el número de posibles agregaciones excede rápidamente cualquier límite factible, de tal forma que se deben seleccionar las agregaciones que mejor satisfagan a los usuarios del negocio.

  Esto no significa que los datos primitivos sean útiles sólo de forma agregada. La retención de los datos originales permite el proceso de búsqueda concreta, ya que se puede requerir examinar los datos que subyacen en los datos agregados seleccionados. Supóngase que después de revisar los totales de ventas mensuales, se requiera ver las ventas totales por punto de venta y por semana y a continuación se centra en las ventas de una sucursal en particular por línea de producto. La retención de los datos originales permite la agregación a cualquier nivel y sobre cualquier dimensión. Esto facilita el proceso de descubrimiento mediante la exploración de la manera más flexible posible.

- **Metadatos**

  La definición final de datos que caracteriza el almacenamiento de los datos es de la de **metadatos**. Literalmente “datos sobre los datos”, los metadatos ofrecen un perfil completo de los elementos de datos, incluyendo su fuente, transformaciones, cualquier resumen, una lista completa de dimensiones, plazo temporal y cualquier información pertinente. Los metadatos también permiten un sistema de clasificación estándar entre los elementos de los datos. Por ejemplo, el término ventas tiene muchos significados en una organización. Para Marketing, una transacción con clientes. Para los gestores de las existencias, es un producto que se tiene que reemplazar. Para el Departamento de Contabilidad, ingresos menos devengos y descuentos. Cada uno de estos conceptos son ventas dentro de un área funcional pero difieren ligeramente en su marco temporal como cuando se producen y accesiones que exigen. Los metadatos permiten una clasificación uniforme que especifica el carácter distintivo de cada elemento de los datos en el Warehouse.

- **Reglas de Negocio**

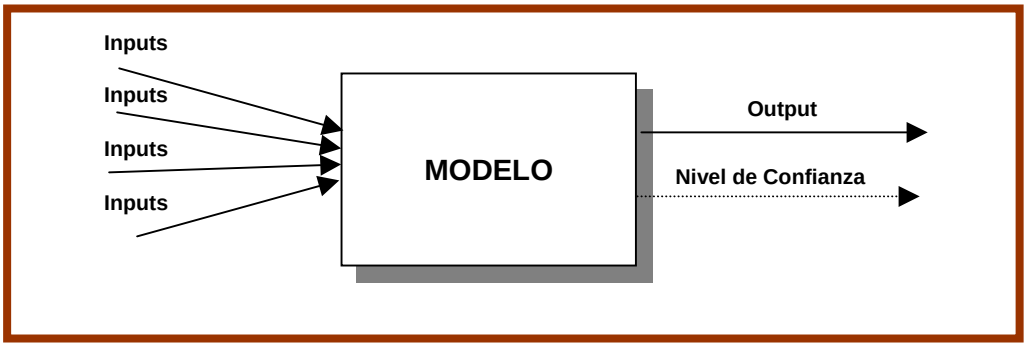
  Finalmente, las reglas de negocio como último nivel de abstracción de los datos. Las reglas de negocio no solo describen la estructura de los datos –este es el rol de los metadatos- sino que describen las relaciones entre los datos y cómo deben aplicarse al negocio. Las reglas de negocio están directamente ligadas al Data Mining ya que técnicas como Canasta de Mercado y Árboles de Decisión, producen reglas explícitas que deben ser aplicadas como reglas de negocio.

♦ **TÉCNICAS DE MODELIZACIÓN Y DESARROLLO DEL MODELO**

Patrones Y Modelos

- **Patrón**
  Se define PATRON a un evento o combinación de eventos que ocurre más veces que lo esperado en los datos analizados.

- **Modelo**
  Un Modelo es una descripción de la base de datos histórica que puede ser aplicado a nuevos datos con finalidad predictiva o descriptiva.



♦ **CONSTRUIR EL MODELO BASADO EN LOS DATOS**

Dependiendo de la hipótesis a testear con el modelo de Data Mining, el analista debe poner en juego todas sus habilidades para construir un modelo a partir de los datos recolectados: seleccionar la herramienta de Data Mining adecuada y el software apropiado (por ejemplo: SAS, S-PLUS, SPSS), los cuales en conjunto con sus conocimientos de estadística, inteligencia artificial y minería de datos, le permitirán arribar al modelo.

El modelo a utilizar depende de los objetivos del Data Mining y de la cantidad y la calidad de los datos disponibles:

- Predicción -> METODOLOGÍA DE DESCUBRIMIENTO DIRECTA
- Descripción -> METODOLOGÍA DE DESCUBRIMIENTO INDIRECTA

Modelo Predictivo	**Modelo Descriptivo**
Explicar el valor de una "variable objetivo" en términos de otras variables. Busca patrones en datos que explican eventos pasados como un camino para predecir eventos futuros	No existe "variable objetivo", simplemente se trata de identificar patrones significativos en los datos
Se trata de estimación, clasificación o predicción de la variable.	Es fundamentalmente aplicado en modelos de segmentación de mercado (clustering) y agrupamiento por afinidad, tales como "canasta de productos"
El objetivo del modelo es "***explicar***" la relación entre los datos.	El objetivo del modelo es "***reconocer***" la relación entre los datos.

Modelo Predictivo	Modelo Descriptivo
Metodología para su aplicación: 1. Identificar fuentes de datos preclasificados 2. Preparar los datos para el análisis 3. Construir el modelo 4. Evaluar el modelo	Metodología para su aplicación: 1. Identificar fuentes de datos 2. Preparar los datos para el análisis 3. Construir el modelo 4. Evaluar el modelo 5. Aplicar el modelo con los nuevos datos 6. Identificar potenciales variables objetivos para la aplicación de un modelo dirigido 7. Generar nuevas hipótesis para testeo

Evaluar que el modelo confirma o rechaza la hipótesis

Finalmente, una vez desarrollado el modelo continúa la etapa de evaluación. Dependiendo de la hipótesis y de la naturaleza del modelo se interpretarán simples valores y/o querys; una colección de asociaciones y reglas generadas por las distintas herramientas o bien, la determinación de la significatividad de la correlación encontrada en modelos de regresión.

Siempre que se pregunte por “si” o “no” a una cuestión dada, el resultado necesita ser evaluado e interpretado, para conocer, por ejemplo, si las diferencias son estadísticamente significativas.

Una apropiada evaluación de los resultados del Data Mining requieren conocimiento analítico y del negocio. Cuando esto no es posible en la misma persona del analista, es necesario la intervención de funciones cruzadas que interpreten y hagan uso correcto de la nueva información.

3-ACTUAR A PARTIR DE LOS RESULTADOS

Esta es la etapa que permite "bajar a la realidad del negocio" los estudios realizados a nivel del Data Mining y que permitirá avanzar sobre la siguiente etapa vinculada con la medición de los resultados. Implica "poner en producción los resultados obtenidos del modelo".

♦ ANÁLISIS DE LOS RESULTADOS

La medición de los resultados provee una realimentación o feedback para mejorar resultados en forma constante. A ésta altura, la Medición de Resultados se refiere específicamente a las medidas de valor comercial que van más allá de las proporciones de respuesta y costos, más allá de los promedios y las desviaciones normales: aquí se trata de medir si el Ciclo de Data Mining ha sido realmente provechoso o "virtuoso".

Aunque la medición del "valor" y la mejora continua se reconoce ampliamente, normalmente se presta menos atención de la que merece. Lo importante aquí es pensar en cada esfuerzo de Data Mining como un pequeño caso de negocios. Comparando las expectativas con los resultados actuales es posible reconocer las oportunidades para explorar la próxima vuelta. ¿Todos los esfuerzos de Data Mining, fueron exitosos para la organización?, ¿Proveen lecciones de aprendizaje para la aplicación de futuros esfuerzos?. La cuestión aquí es identificar qué medir y qué aproximaciones proveen el mejor input para el futuro y ello depende de numerosos factores: las oportunidades de negocio, la sofisticación de la organización, histórica tendencia a los análisis y disponibilidad de los datos.

A título de ejemplo se incluirá el siguiente caso:

Supóngase la medición de una campaña de marketing segmentada
Las mediciones convencionales indican la necesidad de medir tasa de respuesta:

- **¿Cuántos clientes del segmento respondieron a la campaña de marketing?**
  Sin embargo, desde el punto de vista aquí planteado se podrían formular las siguientes cuestiones:

- **El alcance definido en la campaña ha traído a los clientes más rentables?**
  Las organizaciones que poseen un modelo de clientes basado en su rentabilidad pueden responder mejor a esta pregunta.

- **Son estos clientes más fieles que el promedio?**
  El éxito de la campaña debe medirse por los beneficios a largo plazo: clientes que gastan más, refuerzan la relación con la empresa ó que permanezca más tiempo es valioso para el negocio.

- **¿Cuál es el perfil demográfico de los clientes más fieles alcanzados por la campaña?**
  Conocer el perfil demográfico de los clientes permite su aplicación en el futuro sobre prospectos de clientes.

- **La compra del producto generó compras adicionales?**
  Las organizaciones que poseen un modelo de Canasta de Productos están en condiciones de responder más eficientemente a éstas cuestiones.

- Frente a diferentes medios para la realización de campañas (mail físico vs. telemarketing o cupones vs. descuentos , por ejemplo) En cuál de los canales comparados responden los clientes de mayor valor?

Todas estas mediciones impactan en la utilización de esfuerzos futuros: Si una campaña vía telemarketing trae los mejores clientes de la empresa, quizá en el próximo ciclo sea posible comparar los resultados usando diferentes opciones para el canal: quizá leves aproximaciones de venta frente a venta directa. Del mismo modo, el hecho de saber si los clientes atraídos por la campaña son los más rentables en el largo plazo o justo ellos fueron atraídos por un incentivo puntual, permite implementar estrategias de retención basadas en beneficios futuros. Como se ha visto, la respuesta a estas cuestiones subyacen escondidas en los resultados ofrecidos por la campaña.

De las mediciones correctas depende la información que alimentará el siguiente ciclo de Data Mining. Es necesario formular las preguntas adecuadas para colectar la información precisa para las mediciones correctas.

4- MEDIR LA EFECTIVIDAD DEL MODELO DE DATA MINING

Desde el punto de vista de los esfuerzos y recursos que implica la implementación de un proyecto de Data Mining, puede afirmarse que es costoso. El proyecto requiere esfuerzos, tiempo y la aplicación de recursos humanos y tecnológicos para la colección y preparación de los datos, la integración del software, la formulación de los problemas de negocio, la construcción del modelo y el análisis.
Frente a este escenario cómo evaluar el proyecto en relación con la inversión? ¿Cómo evaluar la efectividad del proyecto de Data Mining?

En primer lugar se deberá encontrar la respuesta a las siguientes preguntas:

¿Cuál es el objetivo del ejercicio?

¿En qué medida se ha logrado el objetivo?

¿Se justifica la inversión para el logro de los objetivos?

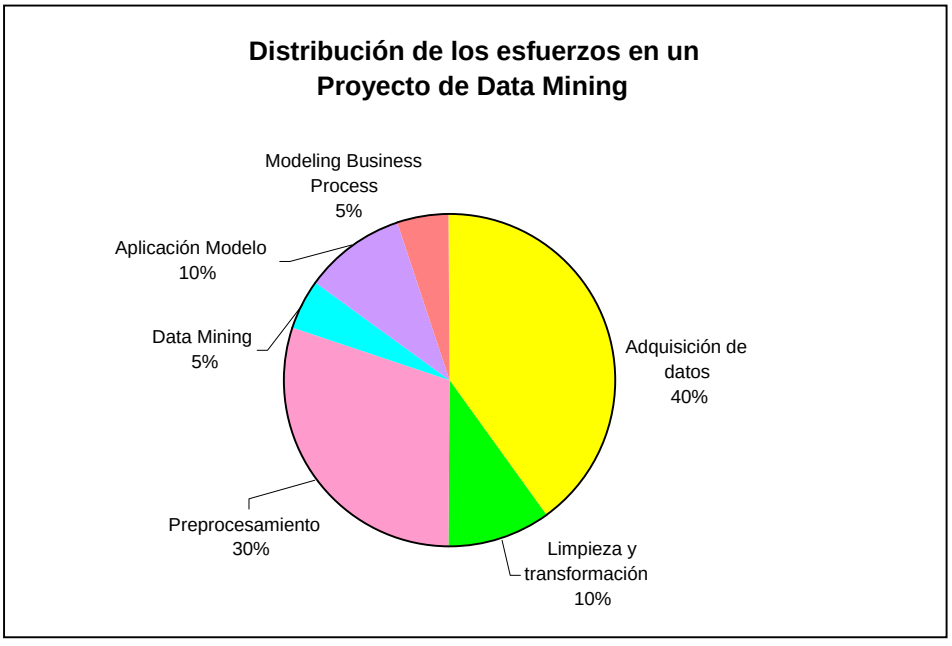
- **LAS MEDIDAS PARA LA EVALUACIÓN DE UN PROYECTO:**

La concordancia de un modelo predictivo con la realidad se mide con relación a la tasa de error, es decir, el porcentaje de casos clasificados o cuya predicción fue incorrecta.

Para ello se dispone de datos de validación y testing sobre los que debe aplicarse periódicamente el modelo de control.

En el caso de los modelos descriptivos, una buena regla es la que proporciona la información más comprensible con la menor "longitud" de expresión de la regla. En definitivas, la medida más importante de efectividad es el retorno de la inversión. La cuestión central aquí es preguntarse si el esfuerzo de Data Mining es necesario para encarar el problema de negocio y en qué medida la respuesta proporcionada brinda soluciones afines.

- **RELACIÓN DE COSTOS EN UN PROYECTO DE DATA MINING**



Desarrollo De Técnicas Para Modelización

El campo de extracción de datos y modelización se acompaña de un amplio número de técnicas analíticas, que van de las más simples a las más complejas. Una característica común a todas, sin embargo, es que la mayoría de las técnicas se basa en principios muy simples que son fácilmente comprensibles.

En la extracción de datos, muchas de las técnicas toman principios básicos tales como ajustarse (reglas de asociación) o analogías biológicas (redes neuronales o algoritmos genéticos) como las bases del procedimiento. Así, el analista de datos puede entender mucho mejor los procedimientos y los resultados. Es apropiado prevenir, sin embargo, contra cualquier noción que en su simplicidad les haga menos "cuantitativos " o rigurosos, en la medida en que cada enfoque puede desarrollar análisis bastantes específicos y detallados.

Visualización

Las técnicas de visualización de datos juegan un papel importante en la extracción de datos, permitiendo al analista de datos emplear uno de nuestros activos más valiosos: la capacidad de la mente humana de procesar reconocer pautas. Tanto si se trata de simples gráficos como de enfoques más sofisticados que crean gráficos multidimensionales, el analista de datos puede obtener un buen conocimiento de las relaciones básicas a través de estos procedimientos.

Programas tales como DIAMOND, desarrollado originariamente por IBM, se ocupa de forma específica de estas técnicas de visualización. Por ejemplo, DIAMOND ha desarrollado procedimientos de presentación de retratos multidimensionales de asociación, parecidos a como se ven en unos gráficos, pero en lugar de representar en sólo dos dimensiones, puede representar hasta nueve dimensiones. El objetivo es representar relaciones de orden superior de la manera más probable para identificar la información oculta que de otra forma se vería oscurecida por los resultados más cuantitativos.

Herramientas para modelización

Muchas de las técnicas multivariantes se emplean en la extracción de datos, como evidencia el fuerte desarrollo de empresas como SPSS y SAS en el campo de la extracción de datos. Por ejemplo, todas las técnicas que se exponen a continuación se emplean extensivamente en la extracción de datos.

- La **regresión múltiple** ofrece un medio directo de confirmación y algunos tipos de exploración de relaciones de dependencia.
- El **análisis factorial** se emplea en la valoración de pautas de variables
- El **análisis cluster** evalúa objetivamente la asociación de objetos entre muchas variables.
- El **análisis discriminante** y la **regresión logística** se emplean a efectos de clasificación.

Consecuentemente, el analista de datos multivariante encontrará un uso extensivo de técnicas familiares en la extracción de datos, residiendo la principal diferencia en la orientación de la investigación.

Reglas de Asociación

Las **reglas de asociación** se parecen mucho al procedimiento de emparejamiento binarios del análisis cluster en que se cuantifica la ocurrencia conjunta de dos eventos. Por ejemplo, considérese estas cuestiones concernientes a dos eventos:

(1) ¿Con qué frecuencia compra cepillos de dientes cuando compra champú?
(2) ¿Con qué frecuencia compra cepillos de dientes cuando compra dentífrico?

Cada uno de estos casos es la base de una regla de asociación, calculada como la ocurrencia conjunta de dos eventos (el porcentaje de veces que se producen). Una regla de asociación no se restringe sólo a dos eventos, ya que podría fácilmente calcularse con qué frecuencia se compras tres o más productos.

Para evaluar las asociaciones se utilizan normalmente dos medidas:

(1) la *Confianza*: medida como la verosimilitud de que se produzca un evento A cuando se produce un evento B
(2) el *Soporte*: que representa el porcentaje de tiempo en que se produce el evento conjunto sobre el total de la población .

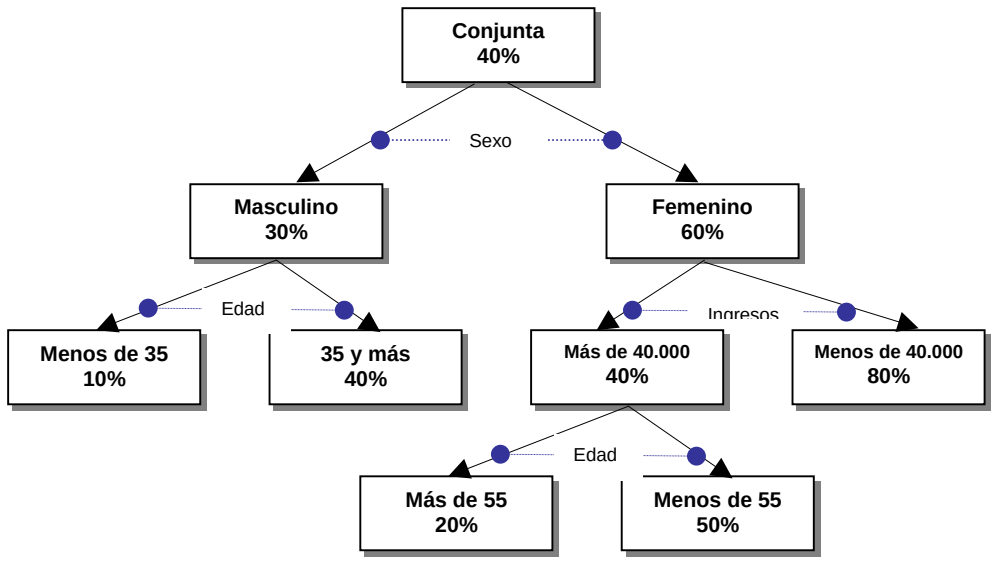
Las reglas de asociación son una herramienta valiosa para ajustar objetos, particularmente en un contexto de mercado, para el cual los objetos son consumidores.

Árboles de decisión

Los árboles de decisión, tienen un aspecto familiar (como el dendograma del análisis cluster), pero se construye y se interpreta de forma completamente distinta.

Los árboles de decisión son particiones del conjunto de datos para maximizar las diferencias de la variable dependiente. Los dos algoritmos más utilizados son CHAID (detector de interacción chi-cuadrado) y CART (árboles de clasificación y regresión). Veamos un ejemplo de partición entre compradores y no compradores mediante tres variables categóricas independientes: sexo, edad y renta.

Un Modelo de Árbol de Decisión



El valor superior de la caja es la etiqueta del grupo, el valor inferior es el porcentaje de compradores. La variable por debajo de un grupo es la utilizada para desglosar ese grupo. Por ejemplo, el sexo se utilizó para desglosar el grupo conjunto

Dado que comprador –no comprador es dicotómico, el objetivo es identificar cuál de las tres variables ofrece la mejor escisión de compradores frente a no compradores (dado que la variable independiente es dicotómica, se considerará sólo el porcentaje de compradores). En conjunto, el 40% de la muestra compra el producto. La mejor escisión del conjunto del grupo es con el sexo, con un 30% de hombres que compran el producto y un 60% de mujeres que compran el producto. A continuación, el procedimiento toma cada uno de esos grupos (hombre y mujeres) y encuentra la variable que mejor escinde cada grupo. Nótese que la misma variable no tiene que escindir cada grupo, en la medida en que una variable podría ser utilizada por hombre y otra variable utilizada por mujeres. Por ejemplo, los hombres se escinden por edad, mientras que las mujeres lo hacen por la renta. El procedimiento continúa hasta que no existen variables independientes sueltas o no existen escisiones significativas pendientes que hacer. El

resultado final es un conjunto de grupos mutuamente excluyentes de clientes que varían en su porcentaje de compra.

Los resultados del ejemplo muestras que el mayor porcentaje de compradores se encuentra en mujeres con una renta por debajo de $40.000. El grupo de menor porcentaje de compradores son los hombres menores de 35 años. Anótese que los grupos que van en segundo lugar en porcentaje bajo y alto respectivamente son ambos del grupo de las mujeres, mientras que las escisiones identificadas ulteriormente identifican un grupo de mujeres con porcentaje más bajo que uno de los grupo de hombres. Cada grupo puede ser mejorado mediante la lectura del diagrama de árbol. Los árboles de decisión ofrecen una forma concisa de desarrollar grupos que son consistentes en sus atributos pero que varían en términos de la variable dependiente.

Redes neuronales

Las **redes neuronales** son una de las herramientas más asociadas con la extracción de datos. Diseñadas después de los trabajos del sistema neuronal del cerebro, las redes neuronales intentan "aprender" mediante ensayos repetidos como organizarse mejor a sí mismas para conseguir maximizar la predicción. Se discutirá este método con más detalle posteriormente, pero se examinará ahora su operativa básica.

El **modelo** se compone de **nodos**, que actúan como **inputs**, **outputs** o **procesadores intermedios**. Cada nodo conecta con el siguiente conjunto de nodos mediante una serie de **trayectorias ponderadas** (parecido a las ponderaciones en un modelo de regresión). Basado en un **paradigma de aprendizaje**, el modelo toma el primer caso, introduce sus datos y a continuación toma una decisión inicial basada en las ponderaciones. Se evalúa el error de predicción y a continuación el modelo hace lo mejor que puede una modificación de las ponderaciones para mejorar la predicción para seguir con el siguiente caso. Este ciclo se repite para cada caso en los que se denomina una **fase de preparación**, cuando se esté calibrando el modelo. Una vez calibrado, el modelo se puede utilizar con otra muestra para evaluar su validez externa.

Muchos analistas de datos ven las redes neuronales como una "caja negra" en la que el analista de datos no controla la estructura del modelo, esto es, cuyos nodos o trayectorias de nodos se conectan por sí mismos. En realidad, todo esto se produce durante el proceso de aprendizaje. Pero esta estructura ofrece una gran flexibilidad. El sistema de red neuronal puede representar relaciones muy complejas, incluso no lineales –algo que es muy difícil de hacer con la mayoría de los modelos multivariantes. Además, en muchos casos puede conseguir una mayor precisión predictiva que el método estadístico comparable. Sin embargo, el analista de datos debe ser precavido en cuanto a la selección de muestras muy específicas y la pérdida de generalidad. Las redes neuronales se han ganado un uso más extendido en las áreas aplicadas que en las académicas, ya que aunque obtengan muy buenos resultados predictivos, lo que es necesario en las aplicaciones, se quedan cortas en las áreas académicas necesitadas de explicación.

Algoritmos Genéticos

El último tipo de técnicas de extracción de datos se denomina también modelo de aprendizaje, pero se basa en una analogía biológica distinta. Los **algoritmos genéticos** copian el proceso evolutivo mediante el uso de la selección natural. Se empieza con un número de soluciones posibles al problema. Las "supervivientes" de esta primera "generación" forman una nueva generación. Algunas tendrán más éxito que la generación anterior y algunas tendrán menos éxito. Los supervivientes en cada generación sucesiva pasan y compiten por la supervivencia en la siguiente generación. Lentamente, a lo largo del tiempo, la selección natural eliminará a las soluciones peores y producirá una mejora global. Este proceso continúa hasta que se consigan unas tasas aceptables de predicción. Al automatizar este proceso, se deben resolver dos asuntos. En primer lugar, debe haber una forma de determinar a los "supervivientes". Existen muchos modos de medir el éxito, pero todos ellos están basados en alguna forma de variable dependiente. En segundo lugar, debe existir un modo para los supervivientes de formar la siguiente generación. Esto se puede conseguir mediante alguna combinación de supervivientes (utilizando una función cruzada) o mutación (una variación aleatoria para cada superviviente).

Una ventaja de los algoritmos genéticos es que generalmente convergen en la solución definitiva: Las desventajas son que esto puede llevar muchas generaciones y se requiere un gran número de individuos. Consecuentemente, este método no es particularmente eficiente, pero los avances en la potencia de cálculo han hecho posible su aplicación en un amplio rango de aplicaciones.

ALGORITMOS Y PROBLEMAS DE APLICACIÓN

Tarea	Algoritmo-Técnica	Ejemplos de problemas de aplicación
Clustering ó Segmentación	Redes neuronales, k-medias, CHAID y otros algoritmos de cluster	Segmentación de mercados
Clasificación	Árboles de decisión, redes neuronales, análisis discriminantes	Descripción del target
Asociación-Dependencia	Redes neuronales, coeficientes de correlación	Análisis de canasta de mercados
Regresión ó predicción	Regresión lineal y no lineal, redes neuronales	Modelos de Precios
Sumarización, reducción de complejidad	ANACOR, HOMALS, factor analysis	Preparación y procesamiento
Análisis de series de tiempo (regresión y detección de cambios)	Modelo ARIMA, redes neuronales	Proyección de venta.-

A continuación se desarrollan en detalle las siguientes técnicas para exploración y modelización que servirán de soporte para el desarrollo del caso de aplicación:

- **Análisis Cluster**
- **Modelos de Asociación o Dependencia**
- **Árboles de Decisión**
- **Redes Neuronales**
- **Análisis de Regresión – Lineal y No lineal**

ANÁLISIS CLUSTER – SEGMENTACIÓN DE MERCADOS

El ***análisis cluster*** agrupa a los individuos y a los objetos en segmentos, de tal forma que los objetos del mismo segmento son más parecidos entre sí que a los objetos de otro segmento. Se intenta maximizar la homogeneidad de los objetos dentro de los segmentos mientras que a la vez se maximiza la heterogeneidad entre los agregados.

¿Qué es el ***análisis cluster***?

El ***análisis cluster*** tiene por objetivo agrupar objetos basándose en las características que poseen. Clasifica objetos (encuestados, productos u otras entidades) de tal forma que cada uno es muy parecido a los que hay en el segmento, con respecto a algún criterio de selección predeterminado. Los segmentos de objetos resultantes deben mostrar un alto grado de homogeneidad interna (dentro del segmento) y un alto grado de heterogeneidad externa (entre segmentos). Por tanto, si la clasificación es acertada, los objetos dentro de los segmentos estarán muy próximos cuando se representen gráficamente y los diferentes grupos estarán más alejados.

En el ***análisis cluster*** el concepto de **valor teórico** es central. El **valor teórico** es el conjunto de variables que representan las características utilizadas para comparar objetos y determina el "carácter" de los mismos. El ***análisis cluster*** es la única técnica multivariante que utiliza el valor teórico especificado por el analista de datos, esto hace crucial la definición que dé el analista de datos al valor teórico del análisis.

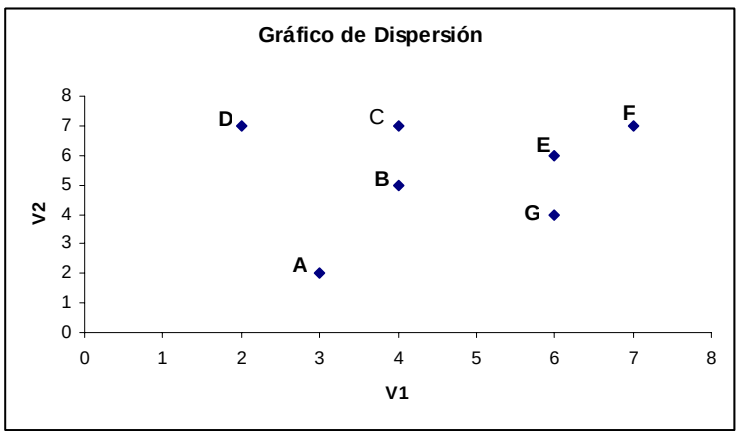
El análisis cluster puede caracterizarse como descriptivo, teórico y no inferencial. No tiene bases estadísticas sobre las cuales deducir inferencias estadísticas para una población a partir de una muestra y se utiliza fundamentalmente como una técnica exploratoria. Las soluciones no son únicas, en la medida en que la pertenencia al segmento para cualquier número de soluciones depende de muchos elementos del procedimiento y se pueden obtener muchas soluciones diferentes variando uno o más de éstos elementos. Además el análisis cluster siempre creará segmentos, a pesar de la existencia de una "auténtica" estructura en los datos. El análisis es totalmente dependiente de las variables utilizadas como base para la medición de similitud. La adición o destrucción de variables relevantes puede tener un impacto substancial sobre la solución resultante.

¿Cómo funciona el análisis cluster?

La naturaleza del análisis cluster puede ilustrarse mediante un simple ejemplo bivariante. Supóngase que se desea determinar los segmentos de mercado en una comunidad reducida basándose en sus pautas de lealtad a marcas y comercios. Se selecciona una reducida muestra de siete encuestados como constrastación de prueba de cómo se aplica el análisis cluster. Se miden dos medidas de lealtad: V1 (lealtad al comercio) y V2 (lealtad a la marca) para cada encuestado y en una escala de 0 a 10. Los valores de cada uno de los siete encuestados se muestran en la **Tabla CL1-Valores de encuestados**, junto con el diagrama de dispersión representando cada observación en las dos variables:

Tabla CL1 - Valores de encuestados

Variable del cluster	Encuestados						
	A	B	C	D	E	F	G
V1	3	4	4	2	6	7	6
V2	2	5	7	7	6	7	4



El objetivo principal del análisis cluster es definir la estructura de datos colocando las observaciones más parecidas en grupos. Para llevar a cabo esta tarea se deben tener en cuenta tres cuestiones básicas:

- ***Medidas de similitud***: Se necesita un método de observaciones simultáneas comparadas sobre dos variables de aglomeración (V1 y V2). Son posibles varios métodos, incluyendo la correlación entre objetos, una medida de asociación utilizada en otras técnicas multivariantes o quizá midiendo su proximidad en un espacio bidimensional de tal forma que la distancia entre las dos observaciones indica similitud.

- ***Formación de segmentos:*** Independientemente de cómo se mida la similitud, el procedimiento debe agrupar aquellas observaciones que son más similares dentro de un segmento. Este procedimiento debe determinar la pertenencia al grupo de cada observación.

- ***Óptima cantidad de segmentos:*** Puede utilizarse cualquier "número de reglas", pero la tarea fundamental es evaluar la similitud media dentro de los segmentos, de tal forma que a medida que la media aumenta, el segmento se hace menos similar. Aquí se enfrenta un tema crítico: pocos segmentos frente a menos homogeneidad. En este caso, se debe buscar un equilibrio entre la definición de las estructuras más básicas (pocos segmentos) que todavía mantienen el necesario nivel de similitud dentro de los segmentos. Una vez que se tienen resueltas estas cuestiones, puede iniciarse el análisis cluster.

MEDICIÓN DE SIMILITUD

Se ilustrará el análisis cluster para las siete observaciones (encuestados A-G) utilizando procedimientos sencillos para cada uno de los asuntos. La similitud será medida de acuerdo a la distancia Euclídea (en línea recta) entre cada par de observaciones. La **Tabla CL2-Matriz de proximidad de distancias euclídeas entre observaciones** contiene las medidas de proximidad entre cada uno de los siete encuestados. Al utilizar la distancia como medida de proximidad, se debe recordar que las distancias más pequeñas indican mayor similitud, de tal forma que las observaciones E y F (1.414) son las más parecidas mientras que A y F (6.403) son las más diferentes.

Tabla CL2-Matriz de proximidad de distancias euclídeas entre observaciones

Observa-ción	Observación						
	A	B	C	D	E	F	G
A	--						
B	3.162	--					
C	5.099	2.000	--				
D	5.099	2.828	2.000	--			
E	5.000	2.236	2.236	4.123	--		
F	6.403	3.606	3.000	5.000	1.414	--	
G	3.606	2.236	3.606	5.000	2.000	3.162	--

FORMACIÓN DE LOS SEGMENTOS

Una vez que se tienen las medidas de similitud, se debe continuar con el procedimiento para la formación de los segmentos. Existen varios métodos, pero para el propósito de este ejemplo se utilizará una regla simple: identificar las dos observaciones más parecidas (cercanas) que no están en el mismo segmento y combinarlas. Se aplica esta regla repetidas veces, comenzando con cada observación en su propio "segmento" y combinando dos segmentos a un tiempo hasta que todas las observaciones estén en un único segmento. A esto se lo denomina ***procedimiento jerárquico*** dado que opera paso a paso para formar un rango completo de soluciones cluster. Es también un ***método aglomerativo*** dado que los segmentos se forman por la combinación de los segmentos existentes.

La **Tabla CL3-Pasos del procedimiento jerárquico** detalla en primer lugar el estado inicial con las siete observaciones en segmentos simples. A continuación se unen los segmentos en el proceso aglomerativo hasta que sólo quede un segmento:

- **Paso 1:** identifica las dos observaciones más cercanas (E y F) y las combina en un segmento, yendo de siete a seis segmentos.
- **Paso 2:** busca los pares de observaciones más cercanos. En este caso, tres pares tienen la misma distancia 2.000 (E-G, C-D y B-C). Se comienza con E-G. G es un miembro único de un segmento, pero E se combinó en el primer paso con F. Así, el segmento formado a este nivel tiene tres miembros: G, E y F.

- **Paso 3:** combina los segmentos de miembro único de C y D
- **Paso 4:** combina B con el segmento de dos miembros C-D que se formó en el paso 3. Hasta este momento se tienen 3 segmentos:
    - Segmento 1: (A)
    - Segmento 2: (B, C y D)
    - Segmento 3: (E, F y G)

La siguiente distancia más pequeña es 2.236 para tres pares de observaciones (E-B, B-G y C-E). Se utiliza sólo una de estas tres distancias, sin embargo, en la medida en que cada par de observaciones contiene un miembro de cada uno de los segmentos existentes (B, C y D frente a E, F y G).

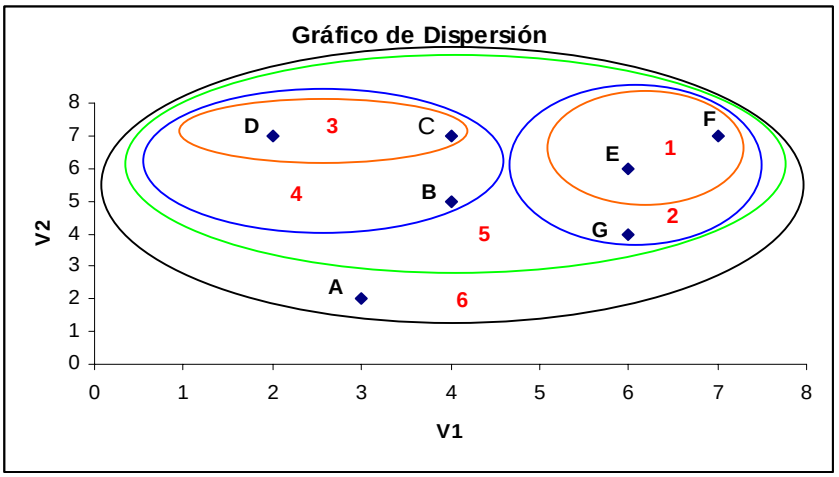
- **Paso 5:** combina los dos segmentos de tres miembros en un único segmento de seis miembros.
- **Paso 6:** combina la observación A con el segmento restante (seis observaciones) en un único segmento a una distancia de 3.162. Se notará que existen tres distancias iguales o menores a 3.162 pero que no se utilizan porque están entre los miembros del mismo segmento.

Tabla CL3-Pasos del procedimiento jerárquico

	Proceso de aglomeración		*Solución Cluster*		
Paso	*Distancia mínima entre observaciones conjuntas (distancia mínima no aglomeradas *)*	*Par de observacio-nes*	*Pertenencia al segmento*	*Número de conglomera-dos*	*Medida de similitud número de segmentos (dentro del segmento)*
	Solución Inicial		*(A) (B) (C) (D) (E) (F) (G)*	*7*	*0*
1	1.414	E-F	*(A) (B) (C) (D) (E-F) (G)*	6	1.414
2	2.000	E-G	*(A) (B) (C) (D) (E-F-G)*	5	2.192
3	2.000	C-D	*(A) (B) (C-D) (E-F-G)*	4	2.144
4	2.000	B-C	*(A) (B-C-D) (E-F-G)*	3	2.234
5	2.236	B-E	*(A) (B-C-D-E-F-G)*	2	2.896
6	3.162	A-B	*(A-B-C-D-E-F-G)*	1	3.420

** Distancia euclídea entre observaciones*

El proceso jerárquico de aglomeración puede graficarse de varias formas. En las figuras a continuación se muestran dos de tales métodos. En primer lugar, dado que el proceso es **jerárquico**, el proceso de aglomeración puede mostrarse como series de agrupaciones anidadas como en el caso de la Figura a). Este proceso, sin embargo, puede representar la proximidad de las observaciones para sólo dos o tres variables de aglomeración del gráfico tridimensional o de dispersión. Una aproximación más habitual es el dendograma, que representa el proceso de aglomeración en un gráfico con forma de árbol. El eje horizontal representa el coeficiente de aglomeración, en este caso la distancia utilizada en la unión de los segmentos. Esta aproximación es particularmente útil en la identificación de los atípicos, como la observación A. También representa el tamaño relativo de los segmentos que varían aunque se hace difícil de manejar cuando aumenta el número de observaciones.



DETERMINACIÓN DEL NÚMERO DE SEGMENTOS EN LA SOLUCIÓN FINAL

Un método jerárquico produce un número de soluciones ***cluster*** –en este caso van de una solución de un segmento a una solución de seis segmentos. ¿Pero cuál se debería elegir? Sabido es que a medida que se aleja de los segmentos de un único miembro, la homogeneidad disminuye. Por otro lado, si se optara por tomar los siete segmentos –que son lo más homogéneos posible- el problema es que no se ha definido ninguna estructura con siete segmentos. En ese caso, se debe ver cada solución cluster a partir de la descripción de su estructura compensada con la homogeneidad de los segmentos.

En este ejemplo, se utiliza una medida muy simple de homogeneidad: las distancias medias de todas las observaciones dentro de los segmentos. En esta solución inicial con siete segmentos, la medida de similitud conjunto es 0 –ninguna observación está emparejada con otra. Para la solución de seis segmentos, la similitud conjunta es la distancia entre las dos observaciones (1.414) unidas en el Paso 1. El Paso 2 forma un segmento de tres miembros (E, F y G), de tal forma que la medida de similitud total es la media de las distancias entre E y F (1.414), E y G (2.000) y F y G (3.162), para una media de 2.192. En el Paso 3, se forma un nuevo segmento de dos miembros con una distancia de 2.000, que provoca que la media conjunta caiga ligeramente hasta 2.144. Se puede proceder a formar nuevos segmentos de esta manera hasta formar una solución de segmento único (Paso 6), en el que la media de todas las distancias de la matriz de distancias es 3.420.

Ahora bien, ¿cómo utilizar esta medida conjunta de similitud para seleccionar una solución cluster? Recuérdese que se está intentando conseguir la estructura más simple posible que represente agrupaciones homogéneas. Si se controla la medida de similitud conjunta a medida que disminuye el número de segmentos, grandes aumentos en la medida conjunta indican que dos segmentos no eran tan similares. En el ejemplo, la medida conjunta aumenta cuando en primer lugar se juntan dos observaciones (Paso 1) y a continuación se hace de nuevo cuando se construye el primer segmento de tres miembros (Paso 2). Pero en los dos pasos siguientes (3 y 4), la medida conjunta no cambia substancialmente. Esto indica que estamos formando otros segmentos prácticamente con la misma homogeneidad de los segmentos existentes. Pero cuando se alcanza el Paso 5, que combina los dos segmentos de tres miembros, se observa un gran aumento. Esto indica que al unir estos dos segmentos se obtiene un único segmento marcadamente menos homogéneo. Considérese entonces la solución ***cluster*** del Paso 4 mucho mejor que la del paso 5. Puede verse también que en el Paso 6 la medida conjunta de nuevo aumenta ligeramente indicando que, incluso aunque la última observación permanezca separada hasta el último paso, cuando se une cambia la homogeneidad del segmento. Sin embargo, dado el perfil bastante aislados de la observación A comparada con el resto, puede ser mejor designar como miembro del ***grupo de entropía***, aquellas observaciones que son atípicos e independientes de los segmentos existentes. Por lo tanto, cuando se revisa el rango de las soluciones ***cluster***, la solución de tres segmentos del Paso 4 parece la más apropiada para una solución definitiva, con dos segmentos de igual tamaño y una única observación atípica.

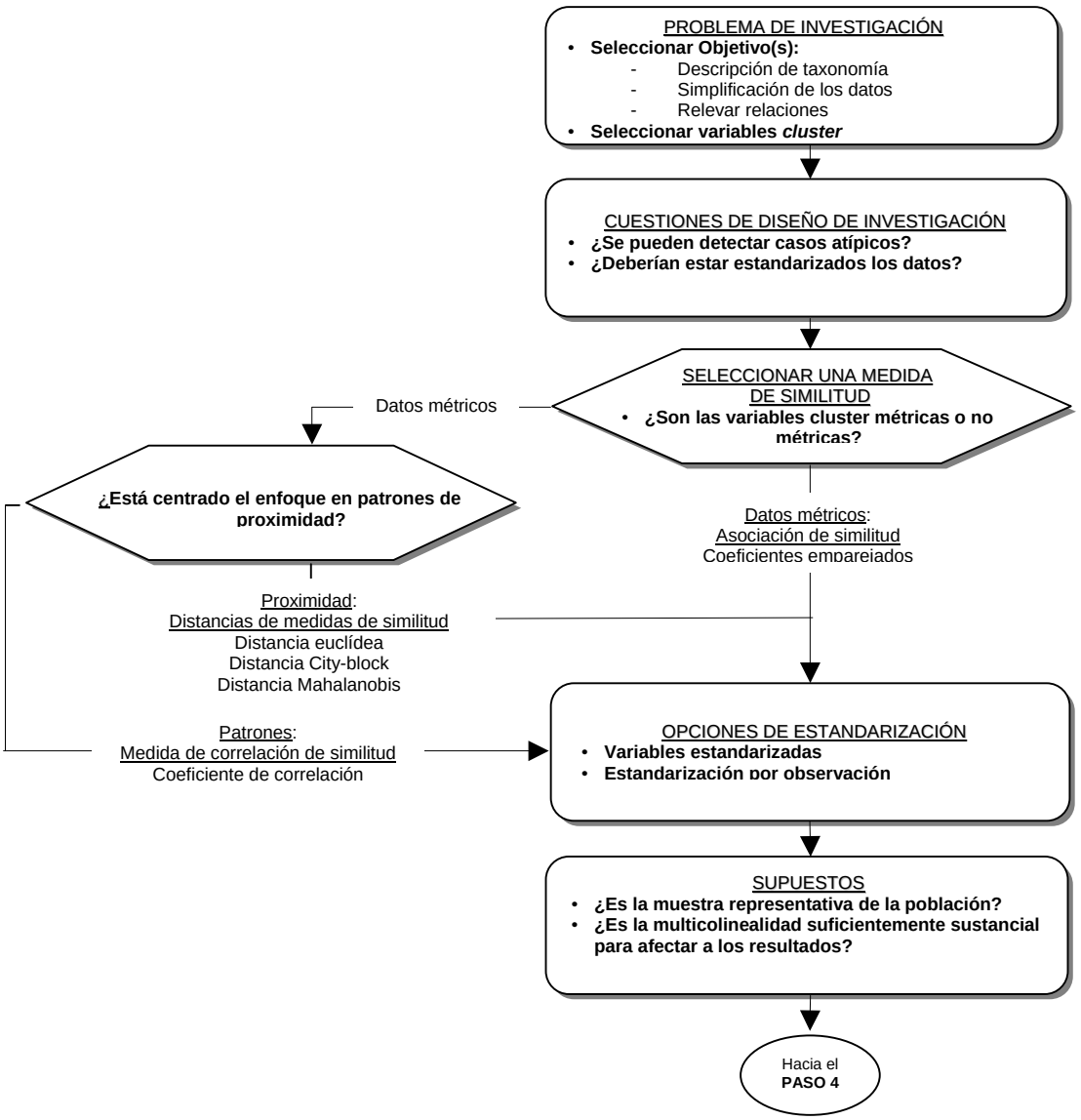
Como podrá verse, en la selección de la solución ***cluster*** definitiva se deja al juicio del observador y es considerado por muchos como un proceso muy subjetivo. Incluso, aunque se han desarrollado métodos más sofisticados para ayudar en la evaluación de las soluciones ***cluster***, sigue recayendo en el analista de datos la decisión final del número de segmentos aceptados en la solución final.

PROCESO DE DECISIÓN CON EL ANÁLISIS CLUSTER

El proceso de decisión consta de los siguientes seis pasos:

1. ***Objetivos del análisis cluster***
2. ***Diseño de investigación mediante análisis cluster***
3. ***Supuestos del análisis cluster***
4. ***Obtención de segmentos y validación del ajuste conjunto***
5. ***Interpretación de los segmentos***
6. ***Validación y perfil de los grupos***

PROCESO DE DECISIÓN EN EL ANÁLISIS CLUSTER



Objetivos del análisis cluster

El objetivo fundamental del análisis cluster es la obtención de un conjunto de objetos en dos o más grupos, basándose en su similitud para un conjunto de características especificadas (valor teórico del análisis cluster). Al formar grupos homogéneos se pueden conseguir los siguientes objetivos:

a) Descripción de una taxonomía
b) Simplificación de datos
c) Identificación de relación

a) ***Descripción de una taxonomía.*** El uso más tradicional del análisis cluster ha sido para propósitos exploratorios y la formulación de una taxonomía –una clasificación de objetos realizada empíricamente. El análisis cluster se ha utilizado para un amplio rango de aplicaciones debido a su capacidad para la partición. Pero el análisis cluster puede generar también hipótesis relacionadas con la estructura de los objetos. Sin embargo, aunque visto principalmente como una técnica de exploración, el análisis cluster puede utilizarse también a efectos confirmatorios.

b) ***Simplificación de datos*****.** En el curso de la obtención de una taxonomía, el análisis cluster también obtiene una perspectiva simplificada de las observaciones. En lugar de ver todas las observaciones como únicas, pueden ser consideradas como miembros de un segmento y perfiladas por sus características generales.

c) ***Identificación de la relación.*** Con los segmentos definidos y la estructura subyacente de los datos representadas en dichos segmentos, el analista de datos tiene un medio de revelar las relaciones entre las observaciones que quizá no fuese posible con las observaciones individuales

Selección de variables del análisis cluster

Cualquiera sea la aplicación, los objetivos del análisis cluster no pueden separarse de la selección de las variables utilizadas para caracterizar los objetos a agrupar. Tanto si el objetivo es exploratorio como confirmatorio, se han restringido efectivamente los resultados posibles por las variables elegidas para el uso. Los segmentos derivados reflejan la estructura inherente de los datos sólo como definida por las variables.

La selección de variables debe hacerse con relación a consideraciones teóricas, conceptuales y prácticas. Cualquier aplicación del análisis cluster debe descansar en cierta lógica en función de la cual se seleccionan las variables.

La técnica del análisis cluster no tiene un medio para diferenciar las variables relevantes de las irrelevantes. La inclusión de una variable irrelevante aumenta la posibilidad de que se creen atípicos sobre estas variables, que pueden tener un efecto importante sobre los resultados. Por tanto, en el análisis nunca se deberían incluir variables indiscriminadamente sino en su lugar elegir las variables utilizando el objetivo de investigación como criterio de selección.

Diseño de investigación mediante análisis cluster

Con los objetivos definidos y las variables seleccionadas, se deben tratar tres cuestiones antes de empezar el proceso de partición:

a) Detección de atípicos
b) Cómo medir la similitud de los objetos
c) ***Estandarización de los datos***

Se pueden utilizar muchos enfoques para responder a éstas cuestiones. Sin embargo, ninguno de ellos ha sido evaluado suficientemente como para ofrecer una respuesta definitiva a cualquiera de estas cuestiones y muchas de éstas aproximaciones ofrecen diferentes resultados para el mismo conjunto de datos. Por tanto, el análisis cluster es más un arte que una ciencia. Por esta razón, se revisarán estos supuestos de forma general ofreciendo una evaluación de las limitaciones prácticas siempre que sea posible.

a) Detección de atípicos

En su búsqueda de una estructura, el análisis cluster es muy sensible a la inclusión de variables irrelevantes. Los atípicos son aquellos objetos con perfiles muy diferentes, la mayoría caracterizados por valores extremos sobre una o más variables. Pero el análisis cluster es también sensible a los atípicos (objetos que son muy diferentes del resto). Los atípicos pueden representar.

(1) observaciones verdaderamente "aberrantes" que no son representativas de la población general, o
(2) Una muestra reducida del grupo (grupos) de la población que provoca una mala representación del grupo (grupos) de la muestra.

En ambos casos los atípicos distorsionan la verdadera estructura y hacen que los segmentos deducidos no sean representativos de la verdadera estructura de la población. Por esta razón, siempre es necesaria una representación preliminar de los atípicos.

Las observaciones identificadas como atípicos pueden evaluarse a efectos de su representatividad respecto de la población y eliminarlos del análisis si se consideran no representativos.

b) Cómo medir la similitud de los objetos

El concepto de similitud es fundamental para el análisis cluster. La **similitud entre objetos** es una medida de correspondencia, o parecido, entre objetos que van a ser agrupados. En primer lugar se establecen las características que definen la similitud. A continuación se combinan las características en una medida de similitud calculada para todos los pares de objetos. De esta forma, cualquier objeto puede ser comparado con otro objeto a través de una medida de similitud. El procedimiento del análisis cluster procede a continuación a agrupar objetos similares juntos en los segmentos.

La similitud entre objetos puede medirse de varias formas, pero tres métodos dominan las aplicaciones del análisis cluster:

b.1.)- Medidas de correlación
b.2.)-Medidas de distancia
b.3.)-Medidas de asociación

b.1.)- Medidas de correlación

El coeficiente de correlación entre las dos columnas de números es la correlación (o similitud) entre los perfiles de los dos objetos. Elevadas correlaciones indican similitud y bajas correlaciones indican falta de ella.

Las correlaciones representan patrones para todas las variables más que las magnitudes. Las medidas de correlación, sin embargo, se utilizan rara vez porque el interés de la mayoría de las aplicaciones del análisis cluster está en las magnitudes de los objetos y no en los patrones de los valores.

b.2.)-Medidas de distancia

Las medidas de similitud de distancia, que representan la similitud como la proximidad de las observaciones respecto a las otras para las variables del valor teórico del análisis cluster, son las medidas de similitud más utilizadas. Las medidas de distancia son en realidad medidas de diferencia, donde los valores elevados indican una menor similitud. La distancia se convierte en medida de similitud utilizando una relación inversa.

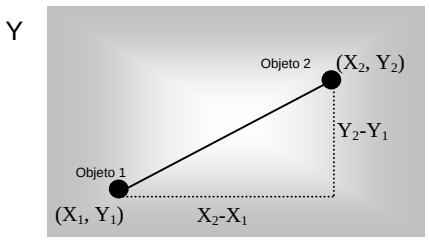
Las medidas de distancia se centran en la magnitud de los valores y representan casos similares que están juntos, pero que tiene pautas muy distintas para todas las variables.

La elección de una medida de correlación en lugar de la medida más tradicional de distancia requiere una interpretación muy diferente. Los segmentos basados en medidas de correlación pueden no tener valores similares en lugar de tener patrones similares. Los segmentos basados en la distancia tienen valores más parecidos para el conjunto de variables, pero los patrones pueden ser bastante diferentes.

Tipos de medida de distancia.

La medida de distancia más utilizada es la **distancia Euclídea**

Ejemplo de Distancia Euclídea entre dos objetos



$$\text{Distancia} = \sqrt{(X_2-X_1)^2 + (Y_2-Y_1)^2}$$

La distancia Euclídea entre los puntos es la longitud de la hipotenusa de un triángulo rectángulo, calculada por la fórmula descripta en el cuadro. Este concepto es fácilmente generalizable para más de dos variables. La distancia Euclídea se utiliza para calcular medidas específicas tales como la simple distancia Euclídea y la distancia Euclídea cuadrada o absoluta, que es la suma de las diferencias al cuadrado, sin tomar la raíz cuadrada. La distancia Euclídea al cuadrado es la medida de distancia recomendada para los métodos de análisis cluster del *centroide* y *Ward*.

Hay varias opciones que no se basan en la distancia Euclídea. Una de las medidas alternativas más utilizadas consiste en reemplazar la diferencia de los cuadrados por la suma de las diferencias absolutas de las variables. Este procedimiento se denomina función de la distancia absoluta. El **enfoque de la distancia absoluta** puede resultar apropiado bajo ciertas circunstancias, pero puede provocar varios problemas. Uno es el supuesto de que las variables no están correlacionadas con el resto; si lo están, los segmentos no son válidos. En la mayoría de los programas del análisis cluster se encuentran otras medidas que emplean variaciones de las diferencias absolutas o las potencias aplicadas a las diferencias (que no sean solo la diferencia de los cuadrados).

Impacto de los valores de los datos no estandarizados.
Un problema al que se enfrentan todas las medidas de distancia es que el uso de datos no estandarizados implica inconsistencias entre las soluciones cluster cuando cambia la escala de las variables. Debería emplearse la estandarización de las variables de aglomeración, siempre que sea conceptualmente posible.

Una medida de distancia Euclídea habitualmente utilizada que incorpora directamente un procedimiento de estandarización es la distancia de Mahalanobis (D^2). El enfoque de Mahalanobis no sólo realiza el proceso de estandarización de los datos a escala en términos de las desviaciones estándar sino que también evalúa la varianza-covarianza unidas dentro del grupo, que ajusta las intercorrelaciones entre las variables. Conjuntos de variables altamente intercorrelacionados del análisis cluster pueden ponderar implícitamente un conjunto de variables en los procedimientos de aglomeración.

Al intentar seleccionar una medida de distancia particular, el analista debe tener presente las siguientes advertencias:

- Diferentes medidas de distancia o un cambio en la escala de las variables pueden llevar a diferentes soluciones cluster.
- Es aconsejable utilizar medidas y comparar los resultados con pautas teóricas o conocidas.
- Cuando las variables están intercorrelacionadas, la medida de distancia de Mahalanobis probablemente sea la más apropiada dado que se ajusta para las correlaciones y ordenaciones de todas las variables igualmente.

b.3.)-Medidas de asociación

Las medidas de asociación de similitud se utilizan para comparar objetos cuyas características se miden solo en términos no métricos (medida nominal y ordinal). Se han desarrollado extensiones de este simple coeficiente de ajuste para acomodar variables nominales de varias categorías o incluso medidas ordinales. Muchos programas informáticos, sin embargo, dan un apoyo limitado a las medidas de asociación y el analista está en muchas ocasiones forzado a calcular en primer lugar las medidas de similitud y a continuación introducir la matriz de similitud en los programas de análisis cluster.

c) ***Estandarización de los datos***

Tipificación de los datos

Con la medida de similitud seleccionada, el analista de datos debe tratar sólo con una cuestión más: ¿deberían tipificarse los datos antes de calcular las similitudes?. Para la respuesta, se deben considerar varios asuntos:

- La mayoría de las distancias medidas son bastante sensibles a las diferentes escalas o magnitudes de las variables.
- En general, las variables con una mayor dispersión (es decir, grandes desviaciones estándar) tienen más impacto en el valor final de similitud.

El analista de datos debe tener cuidado con la ponderación implícita de las variables en función de su dispersión relativa, que sucede con las medidas de distancia.

Estandarización por variables

La forma más común de estandarización es la conversión de cada variable a unas puntuaciones estándar (también conocidas como puntuaciones Z) restando la media y dividiendo por la desviación típica de cada variable. Esta es una opción que muchas veces está incluida en el procedimiento de *análisis cluster*. Es una forma general de una **función de distancia normalizada** que utiliza una media de distancia Euclídea susceptible de transformación normalizada de los datos originales. Este proceso convierte cada puntuación de los datos originales en un valor estandarizado con una media de 0 y desviación estándar de 1. Esta transformación, a cambio, elimina el sesgo introducido por las diferencias en las mediciones de varios atributos o variables utilizadas en el análisis.

Dentro de los beneficios de la estandarización se puede mencionar:

- En primer lugar, es mucho más fácil comparar entre las variables en la medida en que estén en la misma escala. Los valores positivos están por encima de la media y los valores negativos están por debajo; la magnitud representa el número de desviaciones estándar del valor original a partir de la media.

- En segundo lugar, no existe diferencia entre los valores estandarizados cuando sólo cambia la escala.

- Al utilizar las variables estándar se eliminan verdaderamente los efectos debidos a las diferencias de escala no sólo entre las variables, sino también para la misma variable.

- El proceso de estandarización no debería aplicarse sistemáticamente sin considerar sus consecuencias. No hay razón para aceptar absolutamente una solución *cluster* utilizando variables estandarizadas frente a variables no estandarizadas. Si existe alguna relación “natural” reflejada en la escala de las variables, entonces la estandarización puede no ser apropiada.

Estandarización por la observación

Hasta ahora se ha tratado exclusivamente de la estandarización para variables. Si se quiere identificar los grupos de acuerdo a su estilo de respuestas, entonces la estandarización no es apropiada. Pero en la mayoría de los casos lo que se desea es la importancia relativa de una variable con respecto a otra. En este caso, la estandarización por encuestado estandarizaría cada cuestión no por la media de la muestra sino por la puntuación media del encuestado / sujeto. Esta **tipificación entre sujetos** o **tipificación centrada por filas** puede ser bastante efectiva al eliminar efectos de respuesta y es especialmente adecuada para muchas formas de datos de actitud.

Supuestos del análisis cluster

El análisis cluster no es una técnica de inferencia estadística en la que se analizan los parámetros de una muestra en la medida en que puedan ser representativos de una población. Por el contrario, el análisis cluster es una metodología objetiva de cuantificación de las características estructurales de un conjunto de observaciones. Como tal, tiene fuertes propiedades matemáticas pero no fundamentos estadísticos. Los dos asuntos más importantes en que debe centrarse son: la representatividad de la muestra y la multicolinealidad.

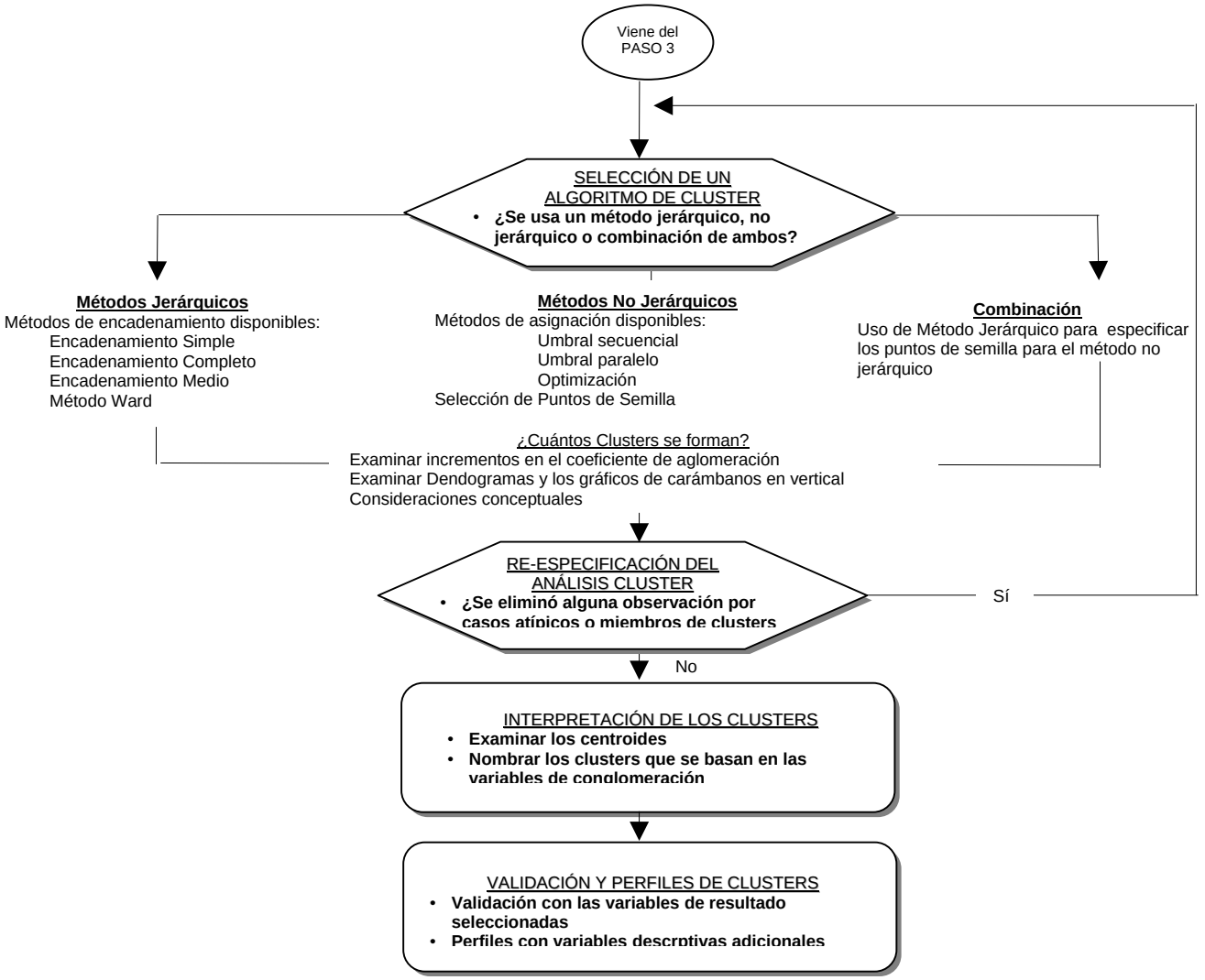
Representatividad de la muestra

La mayoría de los análisis se realizan sobre una muestra de casos proveniente del universo, y los segmentos se derivan en la esperanza de que representen la estructura de la población. La muestra obtenida debe ser realmente representativa de la población. Como ya se ha mencionado, los atípicos pueden ser en realidad producto de una muestra escasa de grupos divergentes que cuando se descartan, introducen sesgos en la estimación de la estructura. El análisis cluster es tan bueno como la representatividad de la muestra. Todos los esfuerzos deben dirigirse a asegurar que la muestra es representativa y que los resultados son generalizables para la población a estudiar.

Impacto de la muticolinealidad

La **muticolinealidad** actúa como un proceso de ponderación no aparente para el observador pero que sin embargo afecta el análisis. En el análisis cluster las variables que son multicolineales están implícitamente ponderadas con más fuerza.
Se debe tratar la muticolinealidad como la discriminabilidad de las variables para llegar a la mejor representación de la estructura.

PROCESO DE DECISIÓN EN EL ANÁLISIS CLUSTER
PARTE II



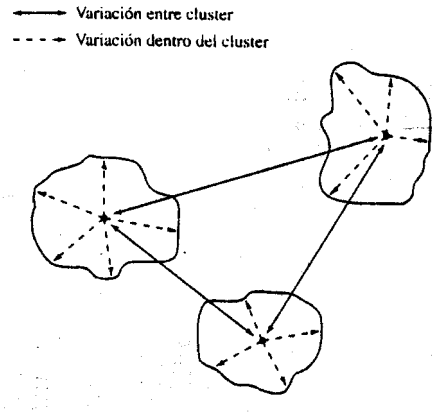
Obtención de segmentos y valoración del ajuste conjunto

Con las variables seleccionadas y la matriz de similitud calculada, comienza el proceso de partición. Se debe seleccionar en primer lugar el **algoritmo de aglomeración** utilizado en la formación de segmentos y a continuación tomar la decisión del **número de segmentos** que se van a formar. Ambas decisiones tienen implicaciones substanciales no sólo sobre los resultados que se obtendrán, sino también sobre la interpretación que se puede derivar de los resultados.

Algoritmo para la obtención de segmentos

El criterio esencial de todos los algoritmos es maximizar las diferencias entre los segmentos relativa a la variación dentro de los segmentos y tal como se muestra en la **Figura-Diagrama de Segmentos que muestra la variación dentro y entre conglomerados.**

Diagrama de Segmentos que muestra la variación dentro y entre conglomerados



Los algoritmos de obtención de los segmentos más utilizados pueden clasificarse en dos categorías generales:

1. ***Algoritmos Jerárquicos***
2. ***Algoritmos no Jerárquicos***

1. Procedimientos Jerárquicos de obtención de segmentos

Los **procedimientos jerárquicos** consisten en la construcción de una estructura en forma de árbol.

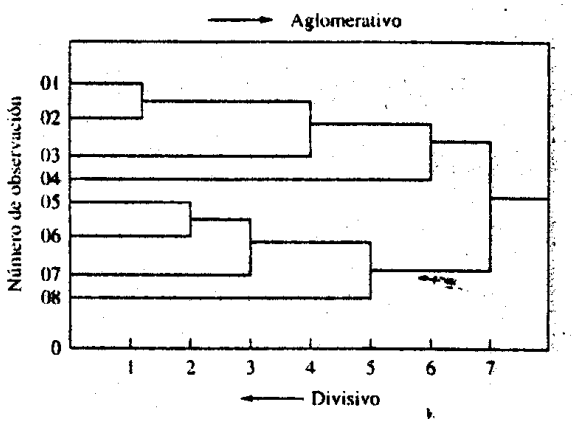
Existen básicamente dos tipos de procedimientos de obtención de segmentos jerárquicos:

1.a)-Métodos de Aglomeración

En los **métodos de aglomeración**, cada objeto u observación empieza dentro de su propio segmento. En etapas ulteriores, los dos segmentos más cercanos (o individuos) se combinan en un nuevo segmento agregado, reduciendo así el número de segmentos paso a paso. En algunos casos, un tercer individuo se une a los dos primeros en un segmento. En otros, dos grupos de individuos formados en un paso anterior pueden unirse en un nuevo segmento. Eventualmente, todos los individuos se agrupan en un único segmento; por esta razón, los procedimientos de aglomeración son denominados como métodos de construcción.

Una característica importante de los procedimientos jerárquicos es que los resultados obtenidos en un paso previo siempre necesitan encajarse dentro de los resultados del siguiente paso, creando algo parecido a un árbol. Por ejemplo, una solución de seis segmentos se obtiene uniendo dos de los segmentos encontrados en el paso de siete segmentos. Dado que los segmentos se forman sólo por unión de los segmentos existentes, se puede rastrear hasta su origen de simple observación cualquier miembro de un segmento. Se muestra este proceso en la **Figura – Dendograma Ilustrativo de la obtención de conglomerados jerárquicos-**, la representación se denomina **dendograma** o gráfico en forma de árbol.

Figura - Dendograma Ilustrativo de la obtención de segmentos jerárquicos-



1.b)-Métodos Divisivos

Cuando el proceso de obtención de segmentos procede en dirección opuesta al método de aglomeración, se denomina **método divisivo**. En los métodos divisivos, se empieza con un gran segmento que contiene todas las observaciones u objetos. En los pasos sucesivos, las observaciones que son más diferentes se dividen y se construyen segmentos más pequeños. Este proceso continúa hasta que cada observación es un segmento en sí mismo. Los métodos aglomerativos que van de izquierda a derecha y los métodos divisivos que van de derecha a izquierda. Dado que los programas informáticos más habituales utilizan los métodos aglomerativos y los métodos divisivos actúan como métodos aglomerativos al revés, se desarrollarán los algoritmos utilizados para los métodos aglomerativos.

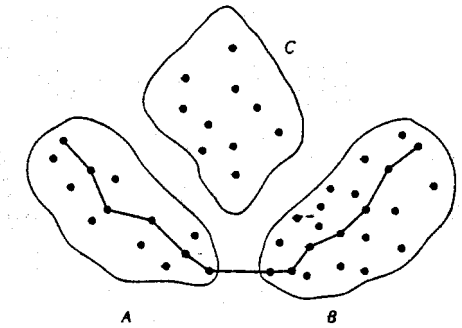
Son cinco los algoritmos más habituales utilizados para desarrollar segmentos, los cuales difieren en la forma en que se calcula la distancia entre segmentos.

i)-Método de encadenamiento simple

El procedimiento de ***encadenamiento simple*** se basa en la distancia mínima. Encuentra dos objetos separados por la distancia más corta y las coloca en el primer segmento. A continuación se encuentra la distancia más corta, y o bien un tercer objeto se une a los dos primeros para formar un segmento o se forma un nuevo segmento de dos miembros. El proceso continúa hasta que todos los objetos se encuentran en un segmento.

La distancia entre dos segmentos cualquiera es la distancia más corta desde cualquier punto en un segmento a cualquier punto en el otro. Dos segmentos se fusionan en cualquier nivel por el vínculo más corto ó más fuerte entre ellos. Los problemas se producen, sin embargo, cuando los segmentos están mal definidos. En tales casos, los procedimientos de encadenamientos simples pueden formar largas y sinuosas cadenas, y eventualmente todos los individuos pueden situarse en una cadena. Los individuos que se encuentran en los límites de una cadena pueden ser muy diferentes.

Figura – Método de Encadenamiento Simple que une los segmentos diferentes A y B

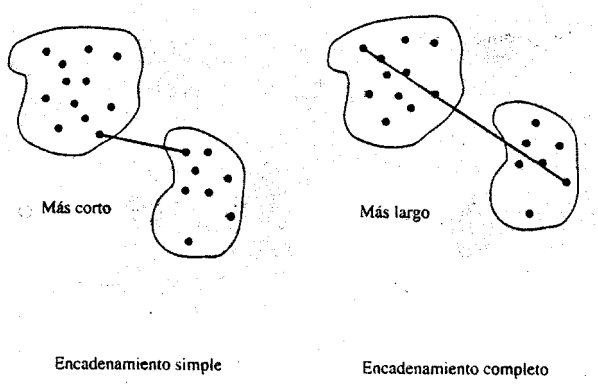


ii)-Método de encadenamiento completo

El procedimiento de ***encadenamiento completo*** es parecido al del encadenamiento simple excepto en que el criterio de aglomeración se basa en la distancia máxima. La distancia máxima entre individuos de cada segmento representa la esfera más reducida (diámetro mínimo) que puede incluir todos los objetos en ambos segmentos. A este método se lo denomina encadenamiento completo porque todos los objetos de un segmento se vinculan con el resto a alguna distancia máxima o por la mínima similitud. Puede decirse que la similitud dentro del grupo es igual al diámetro del grupo. Esta técnica elimina el problema identificado para el encadenamiento simple.

La **Figura – Comparación de las medidas de distancia del encadenamiento simple y completo-** muestra cómo las distancias más cortas (encadenamiento simple) o las más largas (encadenamiento completo) representan la similitud entre grupos. Ambas medidas reflejan solo un aspecto de los datos. El uso de la distancia más corta refleja sólo un único para de objetos (los más cercanos) y el encadenamiento completo también refleja un único par, esta vez los dos más lejanos. Por tanto, es útil ver las medidas como reflejo de la similitud del par de objetos más parecido o del par menos parecido.

Figura – Comparación de las medidas de distancia del encadenamiento simple y completo



iii)-Método de encadenamiento medio

El **método de encadenamiento medio** comienza igual que los métodos de encadenamiento simple o completo, pero el criterio de aglomeración es la distancia media de todos los individuos de un segmento con todos los individuos de otro. Tales técnicas no dependen de los valores extremos, como lo hacen las anteriormente descriptas y la partición se basa en todos los miembros de los segmentos en lugar de un único par de miembros extremos. El enfoque del encadenamiento medio tiende a combinar los segmentos con variaciones reducidas dentro del segmento. También tiende a estar sesgado hacia la producción de segmentos con aproximadamente la misma varianza.

iv)-Método de Ward

En el método de Ward, la distancia entre dos segmentos es la suma de los cuadrados entre dos segmentos sumados para todas las variables. En cada paso del procedimiento de aglomeración, se minimiza la suma de los cuadrados dentro del segmento para todas las particiones (el conjunto completo de segmentos disjuntos o separados) obtenida mediante la combinación de dos segmentos en un paso previo. Este procedimiento tiende a combinar los segmentos con un número reducido de observaciones. También está sesgado hacia la producción de segmentos con aproximadamente el mismo número de observaciones.

v)-Método del centroide

En el método del centroide, la distancia entre dos segmentos es la distancia (normalmente Euclídea simple o cuadrada) entre sus centroides. Los centroides de los grupos son los valores medios de las observaciones de las variables en el valor teórico del segmento. En este método, cada vez que se agrupa a los individuos, se calcula un nuevo centroide. Los centroides de los grupos cambian a medida que se fusionan segmentos. En otras palabras, existe un cambio en un centroide de un grupo cada vez que un nuevo individuo o grupo de individuos se añade al segmento existente.

La ventaja de este método es que se ve menos afectada por los atípicos que otros métodos jerárquicos.

2. Procedimientos no Jerárquicos de obtención de segmentos

En contraste con los métodos jerárquicos, los procedimientos no jerárquicos no implican los procesos de construcción de árboles. En su lugar, asignan los objetos a segmentos una vez que el número de segmentos a formar está especificado. Por tanto, la solución de seis segmentos no es sólo una combinación de dos segmentos desde una solución de siete segmentos, sino que se basa sólo en la búsqueda de la mejor solución de seis segmentos. En un ejemplo simple, el proceso opera de la siguiente manera:

El primer paso es seleccionar una semilla de segmento como centro del segmento inicial, y todos los objetos (individuos) dentro de una distancia umbral previamente especificadas se incluyen dentro del segmento resultante. Entonces se selecciona otra semilla de segmento y la asignación continúa hasta que todos los objetos están asignados. Los objetos pueden entonces asignarse si están cercanos a otro segmento que no sea el original. Existen diferentes aproximaciones para seleccionar las semillas de segmento y asignar objetos. Los procedimientos de aglomeración no jerarquizados se denominan frecuentemente como aglomeración de **K-MEDIAS**, y normalmente utilizan una de las siguiente tres aproximaciones para asignar las observaciones individuales de uno de los segmentos.

i)-Umbral secuencial
ii)-Umbral paralelo
iii)-Optimización

i)-Umbral secuencial

El **método de umbral secuencial** empieza seleccionando una semilla de segmento e incluye todos los objetos que caen dentro de una distancia previamente especificada. Cuando todos los objetos dentro de la distancia están incluidos, se selecciona una segunda semilla de segmento y se incluyen todos los objetos dentro de la distancia previamente especificada. A continuación se selecciona una tercera semilla y el proceso continúa como se ha descrito. Cuando un objeto se incluye en un segmento con una semilla, no se considera a efectos de ulteriores semillas.

ii)-Umbral paralelo

Como contraste, el **método de umbral paralelo** selecciona varias semillas de segmento simultáneamente al principio y asigna objetos dentro de la distancia umbral hasta la semilla más cercana. A medida que el proceso avanza, se puede ajustar las distancias umbral para incluir más o menos objetos en los segmentos. También, en algunas variantes de éste método, los objetos permanecen fuera de los segmentos si están fuera de la distancia previamente especificada desde cualquiera de las semillas de segmento.

iii)-Optimización

El tercer método, denominado **procedimiento de optimización**, es parecido a los otros dos procedimientos no jerárquicos excepto en que permite la reubicación de los objetos. Si en el curso de la asignación de los objetos, un objeto se acerca más a otro segmento que no es el que tiene asignado en este momento, entonces un procedimiento de optimización cambia el objeto al segmento más parecido o cercano.

Selección de puntos de semilla

Los procedimientos no jerárquicos se encuentran disponibles en varios programas informáticos incluyendo los principales programas estadísticos y programas de Data Mining. El procedimiento de umbral secuencial (por ejemplo el programa FASTCLUS en SAS) es un ejemplo de programa de formación de segmentos no jerarquizado diseñado para conjuntos con gran cantidad de datos. Una vez que el analista de datos especifica el número máximo de segmentos permitidos, el procedimiento comienza con la selección de semillas de segmentos, que se utilizan como conjeturas iniciales de las medias de los segmentos. La primera semilla es la primera observación del conjunto de datos sin valores perdidos. La segunda semilla es la siguiente observación completa (sin datos perdidos) que se separa de la primera semilla mediante una distancia mínima especificada. La opción por defecto es una distancia mínima de cero. Una vez que se han seleccionado todas las semillas, el programa asigna cada observación al segmento con las semillas más próximas. El analista de datos puede especificar que los segmentos de semillas se revisen (actualicen) mediante el cálculo de medias de los segmentos de semillas cada vez que se asigna una observación.

Como contraste, los métodos del umbral paralelo (por ejemplo QUICK CLUSTER en SPSS) establecen los puntos de semilla con puntos aportados por el usuario o seleccionado aleatoriamente de las observaciones.

El principal problema a que se enfrentan todos los métodos de formación de segmentos no jerárquicos es cómo seleccionar las semillas de segmento. Por ejemplo, con una opción de umbral secuencial, los resultados del segmento inicial y probablemente del final dependerán del orden de las observaciones en el conjunto de datos y arrastrar el orden de los datos es como afectar a los resultados. La especificación de las semillas de segmento iniciales, como se hace en el procedimiento de umbral secuencial, puede reducir este problema. Pero incluso la selección aleatoria de las semillas de segmento producirá diferentes resultados para cada conjunto de puntos de semillas aleatorios. Por lo tanto, el analista de datos debe ser consciente del impacto del proceso de selección de las semillas de segmentos en los resultados finales.

¿Deben utilizarse los métodos jerárquicos o no jerárquicos?

La respuesta a esta pregunta no es definitiva. En primer lugar, el problema a investigar en el momento puede sugerir un método u otro. En segundo lugar, lo que se aprende con la continua aplicación a un contexto particular puede sugerir un método u otro como el más aconsejable para el contexto, en síntesis: la aplicación de uno u otro método depende del problema a resolver y del contexto de aplicación.

Ventajas y Desventajas de los métodos jerárquicos.

En el pasado, las técnicas jerárquicas de formación de segmentos eran las más populares, siendo el método de Ward y el encadenamiento medio probablemente los mejores disponibles. Los procedimientos jerárquicos tienen la ventaja de ser más rápidos y llevar menos tiempo de cálculo. Pero con el poder de cálculo de hoy en día, incluso las Pc's pueden manejar grandes conjuntos de datos. Los métodos jerárquicos pueden dar una idea equivocada, sin embargo, porque combinaciones iniciales indeseables pueden persistir a lo largo del análisis y llevar a resultados artificiales. De interés específico es el impacto substancial de los atípicos sobre los métodos jerárquicos, particularmente con el método de encadenamiento completo. Para reducir esta posibilidad, el analista de datos puede querer realizar el análisis cluster de los datos repetidas veces, eliminando los atípicos o las observaciones problemáticas. La destrucción de casos, sin embargo, incluso aquellos que no sean atípicos, puede muchas veces distorsionar la solución. También, y más allá de la posibilidad técnica de cálculo, los métodos jerárquicos no son susceptibles de analizar muestras grandes. En ese caso, el analista de datos puede considerar una muestra aleatoria de las observaciones originales para reducir el tamaño de la muestra pero debe cuestionarse cuán representativa es.

Ventajas y Desventajas de los métodos no jerárquicos.

Estos métodos han ganado una creciente aceptación y se aplican cada vez más. Su uso, sin embargo, depende de la capacidad del analista de datos para seleccionar los puntos de semilla de acuerdo a bases prácticas, objetivas y teóricas. Los métodos no jerárquicos tienen varias ventajas respecto de los no jerárquicos:

- *Los resultados son menos susceptibles a los datos atípicos, a la medida de distancia utilizada y a la inclusión de variables irrelevantes o inapropiadas.*

Estos beneficios se obtienen, sin embargo, sólo con el uso de puntos de semilla no aleatorios (es decir, especificados); por lo tanto, el uso de técnicas no jerárquicas con puntos de semilla aleatorios es notablemente inferior a las técnicas jerárquicas. Incluso comenzando con una solución no aleatoria no se garantiza una formación de segmentos óptima de observaciones. De hecho, en muchos casos, el analista de datos obtendrá una solución final diferente para cada conjunto de puntos de semilla especificados. Sólo mediante el análisis y la validación puede el analista de datos seleccionar lo que se considera la "mejor" representación de la estructura, teniendo presentes todas las alternativas que pueden considerarse aceptables.

Una combinación de ambos métodos

Otra aproximación es utilizar ***ambos métodos*** para obtener los beneficios de cada uno. En primer lugar, una técnica jerárquica puede establecer el número de segmentos, los perfiles de los centros de segmentos y la identificación de cualquier atípico obvio. Una vez que se han eliminado los atípicos, las observaciones restantes pueden ser agrupadas mediante un método no jerárquico con los centros de segmentos desde los resultados jerárquicos como los puntos de semilla iniciales. De esta forma, las ventajas de los métodos jerárquicos se complementan con la capacidad de los métodos no jerárquicos para "ajustar" los resultados permitiendo el cambio de pertenencia a un segmento.

¿Cuántos grupos deben formarse?

Quizá el asunto más desconcertante para el analista de datos que utiliza el análisis de segmentos es la determinación del número final de segmentos a formar (también conocida como regla de parada). Desafortunadamente, no existe un procedimiento objetivo o estándar. Dado que no se utiliza un criterio estadístico para la inferencia, los analistas de datos es han desarrollado varios criterios y líneas a seguir para aproximarse al problema. La principal conclusión es que existen procedimientos *ad hoc* que deben ser calculados por el analista de datos, lo que muchas veces implica procedimientos complejos.

- Una clase de reglas de parada que es relativamente simple examina alguna medida de similitud o distancia entre los segmentos a cada paso sucesivo, donde la solución *cluster* se define cuando la medida de similitud excede a un valor especificado o cuando los valores sucesivos entre los pasos dan un salto súbito. Cuando se produce un gran aumento, el analista de datos selecciona la solución cluster previa en la lógica de que su combinación provocó la substancial reducción en su similitud. Se ha mostrado que ofrece decisiones precisas en los estudios empíricos.

- Una segunda clase de reglas de parada intentan aplicar alguna forma de regla estadística o adaptar un test estadístico. Aunque alguno de estos criterios –como el criterio cúbico de selección de segmentos (CCC) que se encuentra en el SAS- han mostrado tener un éxito notable, muchos parecen demasiado complejos para la mejora que ofrecen sobre muestras simples. Se ha propuesto cierto número de procedimientos específicos, pero no se ha encontrado ninguno que sea mejor en todas las situaciones.

También, el analista de datos debería completar el juicio empírico con cualquier conceptualización de las relaciones teóricas que pueda sugerir un número natural de segmentos. Se puede empezar este proceso especificando algún criterio basándose en consideraciones prácticas, como decir, "los resultados serán más manejables y más fáciles de comunicar si se tienen de tres a seis segmentos", por ejemplo, y a continuación resolver para este número de segmentos y seleccionar la mejor alternativa después de evaluar todas ellas. En el análisis final, sin embargo, probablemente sea mejor calcular varias soluciones cluster diferentes y después decidir entre las soluciones alternativas utilizando criterios a priori, juicios prácticos, sentido común o fundamentos teóricos. Las soluciones cluster se verán mejoradas mediante la restricción de la solución de acuerdo con los aspectos conceptuales del problema.

Cuando se identifica una solución aceptable en el análisis cluster, el analista de datos debería examinar la estructura fundamental representada en los segmentos definidos. De interés particular son los tamaños claramente disparatados de los segmentos o segmentos de sólo una o dos observaciones. Los analistas de datos deben examinar los tamaños de los segmentos muy variables desde una perspectiva conceptual, comparando los resultados actuales con las expectativas formadas en los objetivos de la investigación. Más problemáticos son los segmentos de un único miembro, que pueden ser atípicos no detectados en los análisis anteriores. Si aparece un segmento de un único miembro (o de un tamaño muy pequeño respecto de los demás segmentos) se debe decidir si representa un componente estructural válido en la muestra o si debería ser eliminado como no representativo. Si se destruye cualquier

observación, concretamente cuando se emplean soluciones jerárquicas, se debe repetir el análisis cluster y empezar de nuevo el proceso de definición de los segmentos.

Interpretación de los segmentos

El paso de la interpretación implica el examen de cada segmento en términos del valor teórico del segmento o asignar una etiqueta precisa que describa la naturaleza de los segmentos.

Cuando se comienza el proceso de interpretación, una medida utilizada frecuentemente es el centroide del segmento. Si el procedimiento de aglomeración se realizó sobre los datos tal y como se obtuvieron, esto sería una descripción lógica. Si los datos se estandarizaron o si el análisis cluster se realizó utilizando el análisis factorial (de componentes principales), el analista de datos tendría que retroceder a las puntuaciones dadas por los encuestados de las variables originales y calcular los perfiles utilizando estos datos.

Muchas veces el análisis discriminante se aplica para genera puntuaciones de perfiles, pero se debe recordar que las diferencias estadísticamente significativas no indicarían una solución "óptima" porque se esperan diferencias estadísticas, dados los objetivos del análisis del cluster. El examen de perfiles permite una descripción más rica de cada segmento.

Los perfiles y la interpretación de los segmentos consiguen algo más que una descripción. En primer lugar, proporcionan un medio de evaluar la correspondencia de los segmentos derivados de aquellos propuestos por una teoría a priori o por la experiencia práctica. Si se utiliza de forma confirmatoria, los perfiles del análisis cluster ofrecen un medio directo de evaluación de la correspondencia. En segundo lugar, los perfiles de los segmentos ofrecen una vía para realizar evaluaciones de significación práctica. El analista de datos puede exigir que existan diferencias substanciales en un conjunto de variables de elaboración de segmentos u que las soluciones cluster aumenten hasta que surjan tales diferencias. Al evaluar tanto su correspondencia o significación práctica, el analista de datos compara los segmentos derivados con una tipología preconcebida.

Validación y Perfil de los grupos

Dada la naturaleza de alguna forma subjetiva del análisis cluster sobre la selección de una solución cluster "optima", el analista de datos debería tener mucho cuidado en la validación y asegurarse la relevancia práctica de la solución cluster definitiva. Aunque no existe un método único para asegurar la validez y la relevancia práctica, se han propuesto diferentes aproximaciones para ofrecer cierta base a la evaluación realizada por el investigador.

Validación de la solución cluster

La validación incluye los intentos del analista de datos por asegurar que la solución cluster es representativa de la población general y por tanto generalizable a otros objetos y estable en el tiempo. La aproximación más directa en este sentido es realizar análisis cluster para muestras distintas. Esta aproximación, sin embargo, a menudo no es práctica debido a restricciones de tiempos o de costos o a la no disponibilidad de objetos para múltiples análisis cluster. En estos casos, una aproximación común es escindir la muestra en dos grupos. Cada segmento se analiza por separado y se comparan después los resultados. Otras aproximaciones incluyen:

(1) una forma modificada de escisión de muestra por la cual se emplean los centros de segmentos obtenidos desde una solución cluster para definir segmentos a partir de otras observaciones para comparar después los resultados y
(2) Una forma directa de validación cruzada.

El analista de datos también puede intentar establecer alguna forma de **criterio** o **validez predictiva**. Para hacerlo, se selecciona una variable o *variables no utilizadas para formar los segmentos* pero que se sabe que cambian a lo largo de los segmentos. Las variables utilizadas para evaluar la validez predictiva deberían tener un fuerte apoyo teórico o práctico en la medida en que se conviertan en el punto de referencia de selección entre las soluciones de segmento.

Perfiles de la solución cluster

El paso de los perfiles implica la descripción de las características de cada segmento para explicar en qué medida pueden diferir en dimensiones relevantes, lo que implica típicamente el uso del análisis discriminante. El procedimiento comienza una vez que se han identificado los segmentos. El analista de

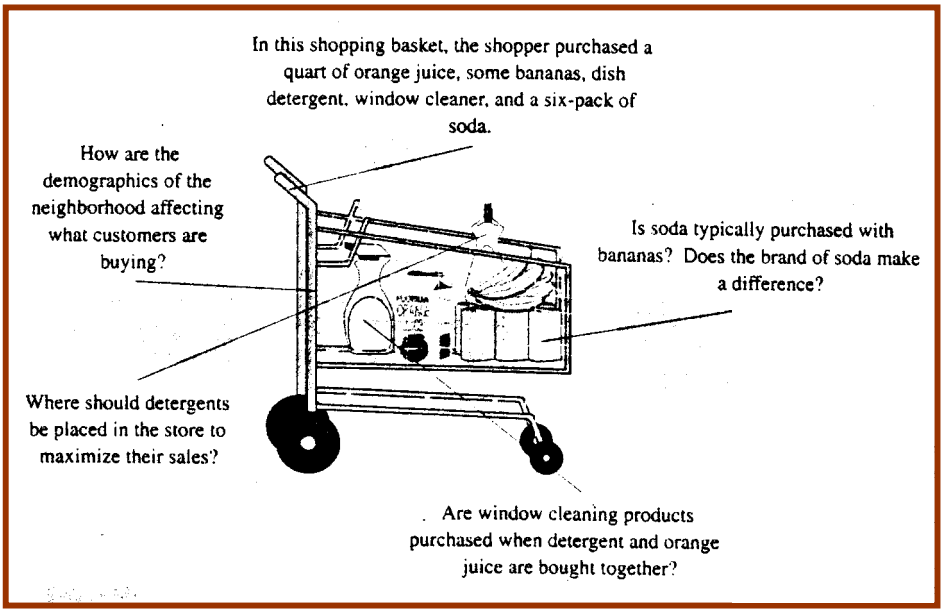
datos utiliza *datos no previamente incluidos* en el procedimiento cluster para perfilar las características de cada segmento. Estos datos son normalmente características demográficas, perfiles psicográficos, pautas de consumo, etc. Aunque puede no existir una razón teórica para que difieran entre los segmentos, tal como requerir la evaluación de la validez predictiva, tienen que tener al menos importancia práctica.

Utilizando el análisis discriminante, el analista de datos compara los perfiles de las puntuaciones medias para los segmentos. La variable categórica dependiente es el conjunto de segmentos previamente identificado y las variables independientes son las demográficas, psicográficas, etc. En resumen, el análisis de perfil se centra en la descripción no de lo que directamente e determinan los segmentos sino de las características de los segmentos una vez que se han identificado. Además, se hace hincapié en las características que difieren significativamente entre los segmentos y aquellas que podrían predecir la pertenencia a un segmento particular.

ANÁLISIS DE CANASTA DE MERCADO – MODELOS DE ASOCIACIÒN Y DEPENDENCIA

Introducción

Para desarrollar esta técnica, considérense las compras de un cliente en cualquier retail. Sabido es que los clientes no son todos iguales y cada cliente compra diferentes sets de productos en diferentes momentos y tiempos.



A partir de la información recolectada de cada compra de clientes, la técnica de Análisis de Canasta de Mercado permite desarrollar modelos que describen el comportamiento de los clientes a partir de cuestiones como: qué compran los clientes, en qué momento y por qué lo hacen. La técnica busca explicar qué productos se compran en forma conjunta en una misma compra, es decir, qué productos son dependientes o asociados entre sí y cuáles son más susceptibles en una promoción. Finalmente se indica con qué intensidad se da el fenómeno encontrado, explicándolo mediante alguna escala numérica.

Esta información es frecuentemente utilizada para definir nuevos circuitos de lay outs; determinar qué productos se compran en forma conjunta e indicar cuándo es conveniente incluir promociones especiales.

La técnica puede ser aplicada en cualquier industria que cuente con la información transaccional de sus clientes. Siempre que los clientes realicen compras de múltiples productos en el mismo tiempo o proximidad, existe potenciales aplicaciones de la misma:

- ***Tarjetas de crédito***: en análisis de transacciones de clientes permite determinar los patrones de compra conjunta o simultánea.
- ***Servicios de telecomunicaciones***: para determinar qué productos se venden en forma conjunta y armar “paquetes de productos” que maximizan la rentabilidad del cliente
- ***Servicios bancarios***
- ***Compañías de seguros***
- ***Servicios medicinales***: las historias médicas de pacientes pueden dar las indicaciones de complicaciones basadas en ciertas combinaciones de tratamientos.

Oportunidad de aplicación de la técnica

La técnica es aplicable cuando se busca describir las dependencias entre variables ya sea a nivel estructural o a nivel cuantitativo especificando la intensidad de la relación mediante alguna escala numérica.

Una regla de asociación sería:

"Si un cliente compra tres líneas de teléfono, entonces comprará también el producto 'llamada en espera'" en el 90% de los casos".

Esta asociación supone un curso de acción sobre los productos, a fin de proponer su venta en forma conjunta y en el mismo paquete:

"tres líneas de teléfono + llamada en espera"

Las reglas de asociación pueden ser fáciles para el entendimiento pero no siempre útiles para su aplicación. A continuación se desarrollarán tres ejemplos de reglas que surgen de datos reales:

- "Los jueves, los clientes que compran pañales compran también cerveza".
- "Los clientes que compran grandes aparatos es probable de contraten acuerdos de mantenimiento"
- "Cuando abre un nuevo negocio de software, uno de los artículos normalmente vendidos es "anillos del retrete""

He aquí tres clásicos ejemplos de los tipos de reglas resultantes de un análisis de canasta de mercado:

- ***Reglas útiles o aplicables***
- ***Reglas triviales***
- ***Reglas inexplicables***

- ***Reglas útiles o aplicables*****:** reglas que contienen buena calidad de información que pueden traducirse en acciones de negocio.

  La regla que trata sobre los pañales y la cerveza sugiere un curso de acción desde el punto de vista del negocio.
  Su explicación radica en que los jueves a la tarde, las parejas jóvenes que adquieren sus compras para el fin de semana se stockean con pañales para sus hijos y cerveza para los maridos.

  Esta regla sugiere un curso de acción respecto del lay out de éstos productos y sus productos complementarios. El curso de acción tendría por objetivo situar los pañales y los restantes productos de bebés cerca de la góndola donde se expone la cerveza, sin olvidar, la proximidad que éstas bebidas deben tener con alimentos como maníes, papas fritas, etc. los cuales deben colocarse también cerca de los productos de bebés.

  De esta manera, se obtiene incrementar las ventas de los productos con mayor margen.

- ***Reglas triviales*****:** reglas ya conocidas en el negocio por su frecuente ocurrencia.

  El segundo ejemplo: "Los clientes que compran grandes aparatos es probable de contraten acuerdos de mantenimiento" responde al perfil de esta regla.
  Esta regla está basada en el análisis transaccional y en la ocurrencia de los casos en repetitivas oportunidades. La oportunidad de negocio en estos casos, es el acople de productos y servicios vendidos en forma conjunta.
  Lo mismo se da en el caso de las líneas telefónicas y el producto "llamada en espera" analizado precedentemente.

- ***Reglas Inexplicables*****:** curiosidades arbitrarias sin aplicación práctica.

  Finalmente el tercer ejemplo: "Cuando abre un nuevo negocio de software, uno de los artículos normalmente vendidos es "anillos del retrete"" responde al perfil de las reglas que promueven curiosidades sin explicación realmente práctica. Muchos otros ítems son vendidos cuando abre un nuevo comercio de este tipo, sin embargo éste en particular se destaca respecto de las cantidades vendidas. Varias investigaciones intentan explicar esta situación: estos productos tienen mayor descuento que otros?; es difícil de encontrar en otro momento?. Sea cual fuere la causa, esta es una curiosidad que surge del análisis de canasta de mercado como expresión creíble.

Cuando se aplica esta técnica, muchos de los resultados obtenidos pueden ser a menudo triviales o inexplicables. Las Reglas Triviales reproducen el conocimiento acerca del negocio. A menudo, los resultados triviales pueden medirse sobre acciones anteriores, como campañas de marketing, pero no mantienen ninguna guía las acciones futuras.

Las reglas Inexplicables, más allá del conocimiento que aportan, no poseen aplicación práctica.

Aplicación de la Técnica

La aplicación de la técnica comienza tomando en cuenta la base transaccional de comercialización de productos o servicios, comúnmente denominados "ítems" a los fines de este propósito de estudio.
Considérese la **Tabla CM1-Transacciones de Puntos de Venta**, que se describe a continuación, la cual ilustra cinco transacciones de cliente que adquieren determinados productos. Dichas transacciones han sido simplificadas, considerando sólo las compras. De qué manera incluir fecha, hora de compra y medio de pago será tratado más adelante. Por ahora, se hará hincapié en las transacciones que proporcionan la información de los productos comprados por los clientes ítems a ítems.

Tabla CM1: Transacciones de Puntos de Venta

Cliente	Ítems
1	Jugo de naranja; soda
2	Leche, Jugo de naranja, limpiador de ventanas
3	Jugo de naranja; detergente
4	Jugo de naranja; detergente; soda
5	Limpiador de ventanas; soda

Usando los datos de la tabla, se puede crear una Tabla de Co-ocurrencia de los ítems que indique el número de veces que cualquier par de productos son comprados en forma conjunta. Surge entonces la **Tabla CM2-Co Ocurrencia de Ítems**:

Tabla CM2-Co Ocurrencia de Ítems

	Jugo de Naranja	Limpiador de ventanas	Leche	Soda	Detergente
Jugo de Naranja	4	1	1	2	1
Limpiador de ventanas	1	2	1	1	0
Leche	1	1	1	0	0
Soda	(2)	1	0	3	1
Detergente	1	0	0	1	2

La Tabla 2 muestra el número de veces que dos productos ó ítems co-ocurren en una misma transacción. Por ejemplo, observando el campo donde la fila "Soda" intercepta la columna de "Jugo de Naranja", se advierte que dos transacciones contienen ambos productos: soda y jugo de naranja. Lógicamente, esto es relativamente sencillo de determinar cuando se tienen, como en este caso, cinco productos y cinco transacciones y donde además, de puede comprobar "a simple vista" que los clientes 1 y 4 han hecho realmente compra de ambos productos.

El valor del campo a lo largo de la fila o columna para un mismo producto representa el número de transacciones que contienen dicho ítem, por ejemplo: Jugo de naranja = 4 tanto sea visto desde la fila y columna de dicho ítem.

La Tabla de Co-ocurrencia de éstos ítems contiene simples patrones de compra, a saber:

- Jugo de Naranja y soda son más probablemente comprados juntos que cualquier otros dos productos.
- Detergente nunca es comprado con Limpiador de Ventanas o Leche.
- Leche nunca es comprada con Soda o Detergente.

Estas simples observaciones son ejemplos de asociación y sugieren una regla formal tal como:

***"Si* un cliente compra Jugo de Naranja, *entonces* el cliente compra también Soda".**

Sin embargo, surge la siguiente pregunta:

¿Cuán buena es esta regla?

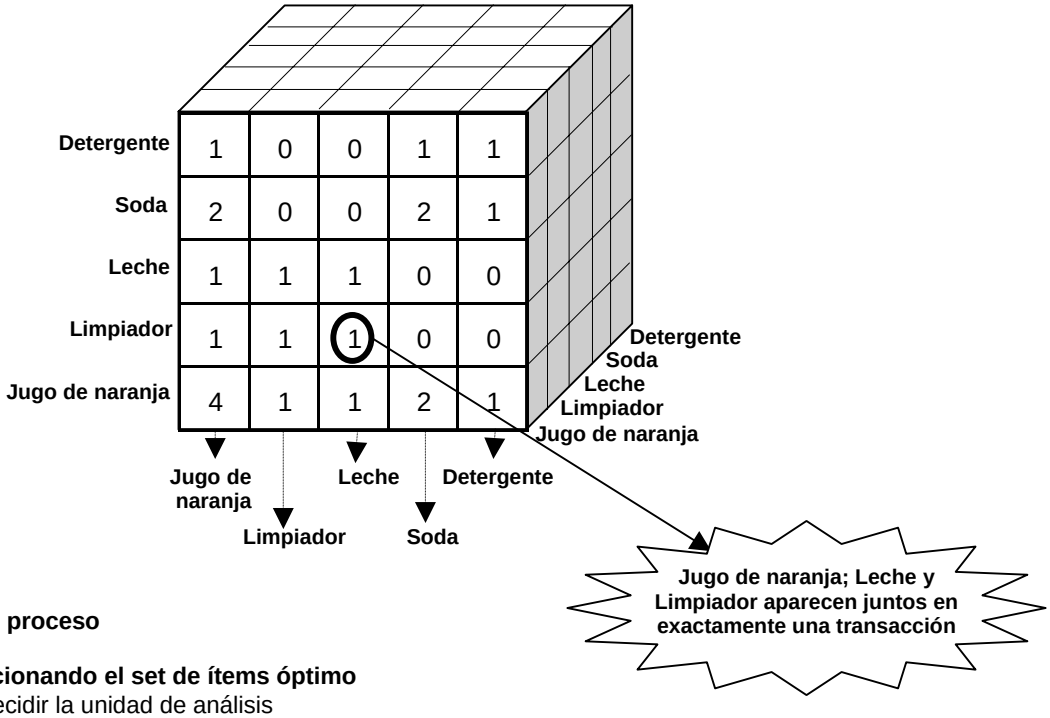
En los datos analizados, se observa que en **dos** de **cinco** transacciones incluye la combinación de productos mencionada: Soda y Jugo de Naranjas. Estas dos transacciones soportan la regla, o bien, presentándolo en porcentaje resulta que esta regla está soportada en el 40% de los casos (2/5).

Puesto que ambas transacciones que contienen soda también contienen Jugo de Naranjas, existe una alto grado de confianza también en la regla, lo cual la justifica como una "buena regla". En efecto, cada transacción que contiene "Soda", también contiene "Jugo de Naranjas", entonces la regla "Si Soda, entonces Jugo de Naranjas" tiene un nivel de confianza del 100%. Si se tomara la regla inversa: "Si Jugo de Naranjas, entonces Soda" puede observarse el menor nivel de confianza existente, ya que de 4 transacciones donde aparece "Jugo de Naranjas" solo en dos de ellas también se compra "Soda". Este nivel de confianza es, entonces de 50%.

El Nivel de Confianza es el valor que indica la cantidad de transacciones (%) que soportan la regla, referida a la cantidad de transacciones que contienen la cláusula condicional.

Soporte es la cantidad de transacciones donde se encuentra la regla.

El ejemplo de Co-ocurrencia presentado mediante la tabla se realiza para combinaciones de dos ítems. Para combinaciones de tres ítems, conviene imaginar un cubo donde se sitúan al margen de cada lado los ítems en cuestión. En estos casos, la interpretación de las combinaciones comienza a dificultarse ya que, con esquemas de cinco productos se tienen 125 combinaciones posibles. En general, el número de combinaciones posibles dado "n" ítems es proporcional al número de ítems elevado a la "n" potencia. La existencia de muchos ítems en el set de análisis complica exponencialmente el tiempo de cálculo.



Etapas del proceso

1. **Seleccionando el set de ítems óptimo**
   - Decidir la unidad de análisis
   - Recopilar los datos
   - Limpiar, reorganizar y armar la tabla para Data Mining

2. **Generar reglas de Co-ocurrencia**
   - Seleccionar el modelo
   - Calcular el modelo

3. **Analizar las probabilidades de determinar las reglas correctas**
   - Interpretar los resultados

A continuación se desarrollarán en detalle estos tres puntos:

1. Seleccionando el set de ítems óptimo

Los datos utilizados para análisis de canasta de productos provienen de transacciones capturadas en el punto de venta. La selección de los ítems para el análisis es una parte crucial del proceso. Esta etapa implica básicamente **decidir la unidad de análisis** en función de las necesidades del negocio.

La existencia de muchos ítems en el set de análisis complica exponencialmente el tiempo de cálculo.
Es importante la construcción de una taxonomía de productos.

Taxonomía de Ítems

En el mundo real, los ítems que poseen códigos de productos o unidades que se enmarcan dentro de categorías jerárquicas son llamados **taxonomías**. En un modelo de canasta de mercado, el punto clave es determinar que nivel de taxonomía es el correcto para usar.
Cuanto más alta es la jerarquía de la taxonomía menor será la cantidad de ítems involucrados en el análisis.
El nivel apropiado de análisis depende de los ítems, de la importancia de producir información con resultados para la acción y de la frecuencia de los datos en el universo de interés.

Calidad de los datos

El dato utilizado para éstos análisis en general requieren de un proceso de limpieza y reorganización para el armado de la tabla final de Data Mining. Al tratarse de transacciones de punto de venta o puntos de contacto con el cliente, están definidos para un propósito operacional como por ejemplo el control de inventarios. Los datos provenientes de sistemas operacionales a menudo requieren procesos de cleaning y transformación, antes de convertirse en una fuente de datos para el análisis. Si el análisis se alimenta de múltiples fuentes de datos es necesario resolver cuestiones como formatos, codificación, posibilidad de extracción, contenido de las variables, etc. antes de su aplicación en el modelo. Esta cuestión es particularmente crítica en un análisis de canasta de mercado ya que éste tipo de análisis depende de la sumarización de transacciones de punto de venta.

2. Generar reglas de co-ocurrencia

Calcular el número de veces que una combinación de ítems aparece en una transacción de datos es posible, pero una combinación de ítems no es una regla.

Una regla es una estructura enunciativa que contiene una condición y un resultado.

Usualmente las reglas se representan de la siguiente manera:

Si *Condición* entonces *Resultado*

Construcciones como la tabla de **Co-Ocurrencia** proveen la información acerca de las combinaciones de ítems que ocurren más frecuentemente en las transacciones.

A título de ejemplo se consideraron las combinaciones de 3 ítems: ‘A’, ‘B’, ‘C’. y las reglas donde aparecen éstos 3 ítems en la regla y exactamente uno de ellos en el resultado:

- ***Si ‘A’ y ‘B’, entonces ‘C’***
- ***Si ‘A’ y ‘C’, entonces ‘B’***
- ***Si ‘B’ y ‘C’, entonces ‘A’***

La **Tabla CM3-Probabilidades de tres ítems y su combinación**; provee muestra las distintas probabilidades y combinaciones de los ítems.
Como éstas tres reglas contienen los mismos ítems, ellas tienen los mismos soportes en los datos, 5%. Ahora bien, cuál es su nivel de confianza?

Tabla CM3-Probabilidades de tres ítems y su combinación

'A'	45%
'B'	42.5%
'B'	40%
'A' y 'B'	25%
'A' y 'C'	20%
'B' y 'C'	15%
'A' y 'B' y 'C'	5%

El nivel de confianza para éstos tres ítems se muestra en la **Tabla CM4-Nivel de Confianza de las reglas**. Allí se indica que la regla "*Si 'B' y 'C' entonces 'A'*" tiene un nivel de confianza del 33%, y es equivalente decir que cuando B y C aparecen en una transacción, eso indica un 33% de chance que 'A' también aparezca en ella. Esto es, una de cada tres veces 'A' ocurre con 'B' y 'C' mientras que en las otras dos veces 'A' no ocurre.

Tabla CM4-Nivel de Confianza de las reglas

Regla	P(condición)	P (condición y resultado)	Nivel de confianza
Si 'A' y 'B' entonces 'C'	25%	5%	0.2
Si 'A' y 'C' entonces 'B'	20%	5%	0.25
Si 'B' y 'C' entonces 'A'	15%	5%	0.33

La regla que posee el mayor nivel de confianza es la mejor regla, en consecuencia se seleccionará "*Si 'B' y 'C' entonces 'A'*". Sin embargo, esta regla es peor que decir que 'A' aparece en la transacción en forma aleatoria. 'A' ocurre en el 45% de las transacciones mientras que la regla proporciona el 33% de confianza. La regla es más errónea que la aparición aleatoria de la variable.

Esto sugiere otra medida llamada "***Improvement***" (Mejora). "Improvement" indica la capacidad predictiva de la regla: cuánto mejor una regla está prediciendo el resultado que asumir el resultado de primera mano.

$$\textit{Improvement} = \frac{p\ (\text{condición y resultado})}{P\ (\text{condición}) * p(\text{resultado})}$$

Cuando:

Improvement >= 1; la regla tiene valor predictivo
Improvement < 1 ; la regla no tiene valor predictivo

En la **Tabla CM5-Improvement para cuatro reglas**, se muestra el valor de éste parámetro. Como se puede observar, la mejor regla es la que posee 2 ítems. La regla "Si 'A' entonces 'B'" tiene un Improvement de 1.31 y es precisamente 1.31 veces mejor predictora que el azar.

Tabla CM5-Improvement para cuatro reglas

Regla	SOPORTE	CONFIANZA	IMPROVEMENT
Si 'A' y 'B' entonces 'C'	5%	0.20	0.50
Si 'A' y 'C' entonces 'B'	5%	0.25	0.59
Si 'B' y 'C' entonces 'A'	5%	0.33	0.74
Si 'A' entonces 'B'	25%	0.59	1.31

Reglas de Disociación

De la misma forma que las Reglas de Asociación existen las Reglas de Disociación. La estructura de ambas es similar con la excepción de que las reglas de disociación poseen el conector "y no" en lugar al conector "y".

Si 'A' y no 'B' entonces 'C'

ÁRBOLES DE DECISIÓN – REGLAS DE CLASIFICACIÒN

Los árboles de decisión son herramientas poderosas para representar reglas de clasificación de individuos en grupos disjuntos.

Los árboles de decisión se aplican en caso de contar con una variable objetivo, por ejemplo, la tasa de respuesta a un mailing, y cuyo comportamiento se requiere explicar. Esta variable es la que orienta el aprendizaje del modelo y por ese motivo se define a estas técnicas como de "aprendizaje supervisado". La segmentación intentará separar a los individuos en grupos homogéneos según la variable objetivo.

Existen dos tipos de Árboles de Decisión:

- **Árboles de Clasificación**: asigna los registros en clases ó grupos y provee un nivel de confianza relativo a la clasificación.

- **Árboles de Regresión**: utilizados para predecir el valor de una variable que asume valores numéricos.

El atractivo mayor de la técnica de árboles de decisión es que, en oposición a las redes neuronales, los árboles representan reglas, y en la mayoría de las aplicaciones prácticas esto es lo que más importa.

La técnica se aplica sobre la base de algoritmos que permiten la construcción de los modelos. Existen una variedad de algoritmos aunque los más utilizados son: CART -Clasification and Regression Trees- y CHAID –Chi Squared Automatic Interaction Detection-. Otros algoritmos como C 4.5 han ganado popularidad ya que están disponibles en los softwares de Data Mining.

Todos los árboles poseen una estructura similar. Su construcción comienza a partir del nodo raíz y se extiende hasta los nodos finales. Los datos ingresan a través del nodo raíz y fluyen a través del árbol, como consecuencia de la aplicación de pruebas ó tests que determinan los siguientes nodos. Existen diferentes algoritmos de aplicación para seleccionar las pruebas que mejor discriminen las clases del target. Este proceso es repetitivo hasta que los datos arriben al nodo final. Una vez que todos los datos arriban a este punto, se obtienen como resultado "caminos" o "reglas" que explican el comportamiento de las variables iniciales (TARGET) con relación a otras (PREDICTORAS).

Como se ha dicho anteriormente, la mayor oportunidad de la técnica es la generación de reglas. Una vez desarrollado el árbol, se debe analizar cada "camino" en particular, teniendo en cuenta que cada uno de ellos representa una "regla" y algunas reglas son mejores que otras.

En cada uno de los nodos, se puede medir:

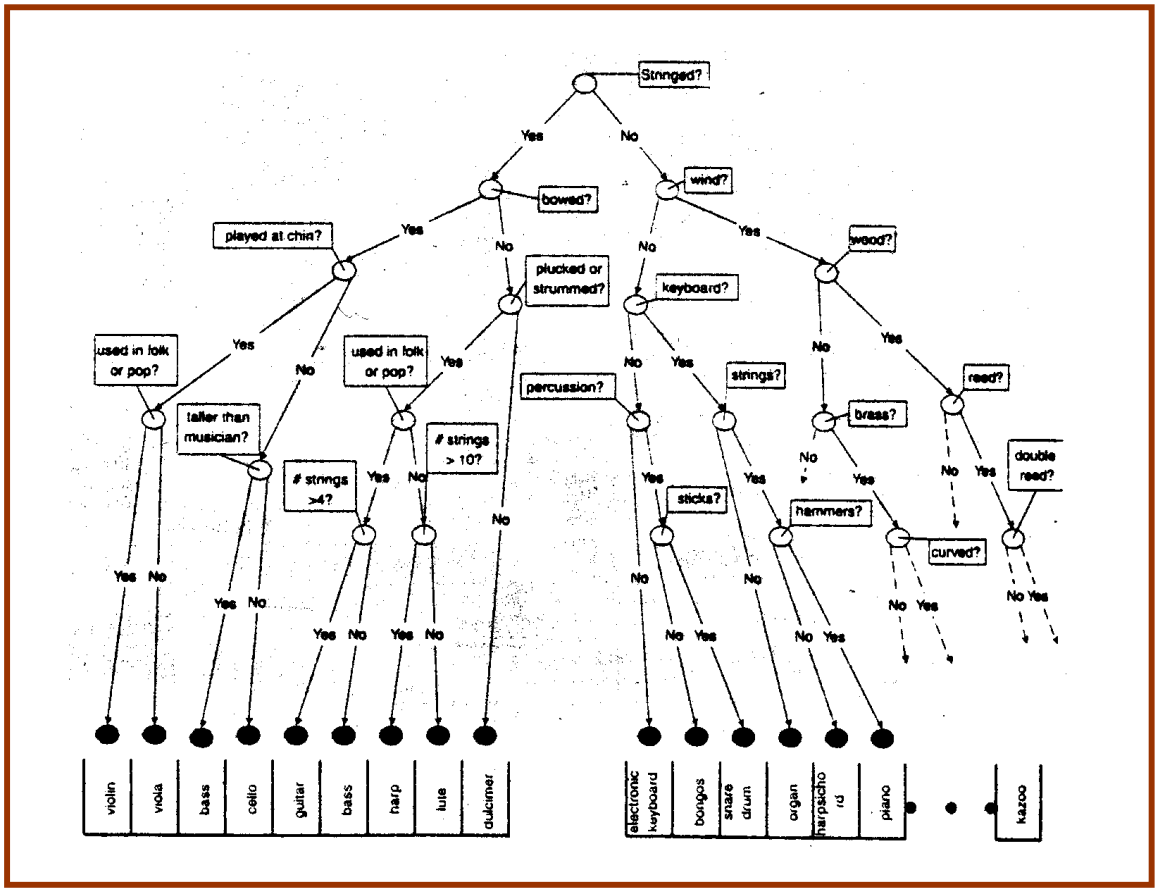
- **El número registros que incluye**
- **Clasificación de los archivos como si fuera el nodo final**
- **El porcentaje de registros clasificados correctamente**

Los algoritmos que construyen árboles de decisión comienzan con la búsqueda de la prueba o test (predictor) que mejor fraccione o clasifique los datos entre las categorías deseadas. La prueba realizada sobre el nodo raíz genera dos o más subdivisiones. En cada nivel subsiguiente del árbol, los subconjuntos resultantes son nuevamente divididos según las reglas o predictores que mejor discriminen la clasificación. En lo sucesivo, el árbol continúa creciendo hasta que ninguna otra subdivisión sea posible encontrar para los datos.

Los árboles de decisión crean cajas o boxes para los datos

Antes de entrar en detalle acerca de la aplicación de los algoritmos CART, CHAID y C4.5 y la técnica, es importante destacar que el resultado de los árboles se representa a través de diagrama de cajas que describen las reglas de clasificación. El resultado final de éstos diagramas de cajas son gráficos de una dimensión. El test o prueba asociado al nodo raíz de un árbol de decisión divide la línea en dos o más secciones. Cada sección es a su vez subdividida a lo largo del árbol.

En la figura que se muestra a continuación, por ejemplo, el salto en el centro de la fila de cajas corresponde al campo analizado en el nodo de raíz: “Todos los instrumentos de cuerda quedan del lado de la rama izquierda.



Representación en varias dimensiones

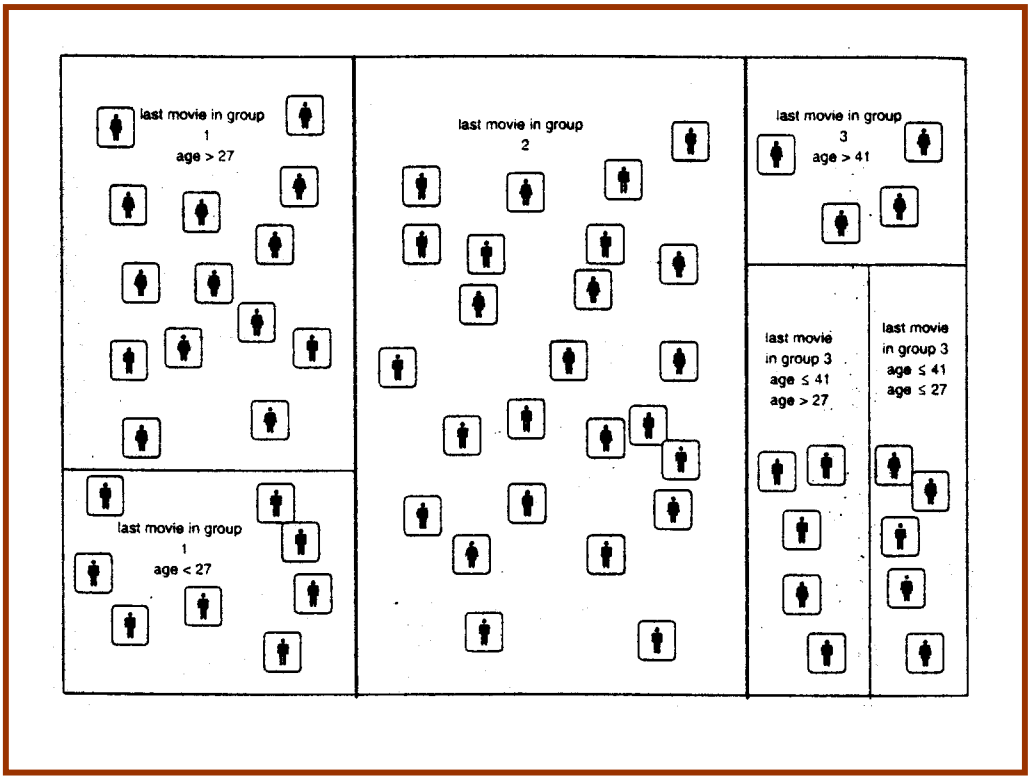
Los gráficos de dos dimensiones X e Y son los más familiares para explicar la relación entre dos variable. En la estructura de diagramas de cajas, se aprovecha de las habilidades de un esquema bi-dimensional para representar un árbol de decisión y los registros son clasificados como una clase de colecciones anidadas en dos dimensiones.

En el nodo inicial del árbol, la división es realizada sobre algún campo en particular del dato. En el extremo de la caja de un diagrama de cajas, la axisa horizontal representa ese campo. Se divide el extremo de la caja en secciones, una para cada nodo al siguiente nivel del árbol. Si la división del campo representa una variable categórica como “animal, vegetal o mineral” generalmente se selecciona la posición en la división tal que el tamaño de cada sección sea proporcional al número de registros que caen en él. Si el campo contiene una variable continua, se tiene una opción adicional considerando la línea horizontal en el eje de las X del gráfico y ubicando en la línea vertical valores usados de acuerdo a la función de división. En cualquier caso, la axisa vertical para cada caja representa el campo usado como divisor o discriminante para cada nodo. En general, existirán campos diferentes para cada caja.

Ahora se tiene un nuevo set de cajas, cada uno de los cuales representa un nodo en el tercer nivel del árbol. Y así podrá continuarse dividiendo las cajas hasta llegar al nivel máximo del árbol. Mientras que los árboles de decisión posean agrupamientos no uniformes, algunas cajas podrán ser subdivididas más a menudo que otras. El diagrama de cajas permite representar reglas de clasificación que dependen de cualquier número de variables sobre un gráfico de dos dimensiones.

El resultado del diagrama es muy expresivo, ya que muestra las reglas de clasificación para los datos en análisis.

Visto de esta manera, es natural pensar que los árboles de decisión constituyen una manera de "armar cajas" alrededor de grupos de similares características. Todas estas características son clasificadas de la misma manera porque todas ellas comparten la regla que define la caja. En contraste con los métodos de clasificación estadísticos clásicos como lineales, logísticos y discriminantes cuadráticos que intentan dividir datos en las clases, dibujando una línea o la curva elíptica a través del espacio de los datos. Ésta es una distinción fundamental: acercamientos estadísticos que usan una sola línea para encontrar el límite entre las clases son débiles cuando hay varias maneras muy diferentes para un registro formar parte de una clase determinada.



Algoritmos frecuentemente utilizados

CART	C4.5	CHAID
El algoritmo de CART es uno de los más populares para la construcción de árboles de decisión.	Es el más reciente algoritmo para la generación de árboles de decisión	Es el algoritmo más distribuido a través de los software de Data Mining.
Publicado en 1984 por L.Briemen & Asoc.	Publicado por Australian researcher, J.Ross Quinlan.	El primer algoritmo publicado por J.A.Hartigan en 1975
Muchas de las herramientas de reglas de inducción disponibles comercialmente usan alguna variante de este algoritmo para generar sus reglas.	Utilizado en Clementine Data Mining software: Integral solutions, LTd.Clementine. Bajo un nombre diferente también es comercializado por NCR.	Utilizado en software comercializados por SPSS y SAS.
Genera árboles con divisiones binarias	Produce árboles con variables números de divisiones	La gran diferencia entre éste algoritmo y los anteriores es que CHAID controla el crecimiento del árbol, es restrictivo a variables categóricas y las variables continuas deben ser cortadas en rangos o reemplazadas por clases como "alto", "media" y "baja".

Construcción de los Árboles de Decisión

Un árbol de decisión se construye sobre la base de un proceso conocido como "Particionamiento Sucesivo". El mismo, es un proceso interactivo de división de los datos en sucesivas partes. Inicialmente, todos los registros –cuya preclasificación determina la estructura del árbol- son incorporados en una gran y única caja. El algoritmo aplicado trata de particionar los datos, usando todas las posibles divisiones binarias sobre cualquier campo. Por ejemplo: si se tomaron 72 valores que corresponden a edades entre 18 y 90 años, el algoritmo entonces considera una división entre los valores 18 y más de 1; otra posibilidad es particionar entre 18 ó 19 y 20 o más por otro lado y así sucesivamente. El algoritmo luego selecciona la división que particiona el dato en dos partes que son más puras que el original. Esta división o proceso de particionamiento es aplicado entonces a cada uno de las nuevas cajas. El proceso continúa hasta que no se encuentren más divisiones útiles. El corazón del algoritmo es la regla que determina la separación inicial.

Encontrando la División Inicial

El proceso se inicia con un entrenamiento consistente en la preclasificación de los registros. Preclasificación significa que el campo target ó *variable dependiente*, tiene una clase conocida. El objetivo es construir un árbol que distinga alrededor de las clases. Esto significa que el árbol puede usarse para asignar una clase al campo target de un nuevo registro basado en valores de otros registros o *variables independientes*.

Para simplicidad, se asume que sólo se tienen dos clases de target y que cada una de las divisiones es un particionamiento binario. El criterio de división generaliza fácilmente a las clases múltiples y permite la formación de caminos o reglas a través de sucesivas particiones binarias.

El primer paso es decidir cual es la variable independiente que da lugar a la mejor división. La mejor división se define como aquella que mejor realiza la separación de registros en grupos donde solo predomina una clase. La medida utilizada para evaluar un divisor potencial es la reducción en *diversidad (otra forma de decirlo es "el incremento de la pureza"*. Porque el concepto de "diversidad" (o conservación de la pureza) es el verdadero corazón del método del árbol de decisión.

A continuación se presenta un breve ejemplo que permite calcular el índice de diversidad en un conjunto de datos disjuntos:

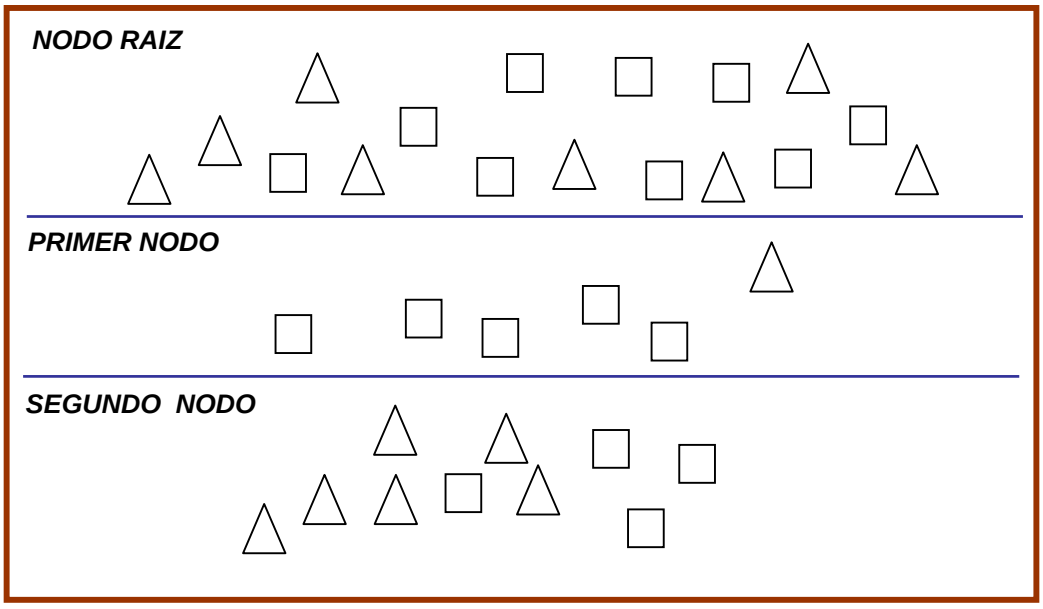


Figura_ Una población diversa es dividida en dos sub-poblaciones con mayor nivel de pureza

Resulta intuitivamente obvio en el gráfico adjunto que el nodo raíz tiene un mayor índice de diversidad que el resto de los nodos. El índice de diversidad es la probabilidad que el segundo elemento dentro de un grupo pertenezca a una clase diferente que el primero.
El nodo raíz del árbol contiene 16 elementos: 9 cuadrados y 7 triángulos.
La P de encontrar cuadrados en el primer nodo es = $P_{(cuadrados)}$ = 9/16 = 0.56
La P de encontrar triángulos en el primer nodo es = $P_{(triángulos)}$ = 7/16 = 0.44

Para calcular la probabilidad de encontrar dos diferentes figuras en segundos intentos, es necesario considerar la probabilidad de encontrar ejemplos de las mismas figuras en el primer intento. Luego, la P del segundo intento es igual a la P^2 del primer intento
*La P de encontrar cuadrados dos veces seguidas = 0.56*0.56*
*La P de encontrar triángulos dos veces seguidas = 0.44 *.0.44*
El índice de diversidad es 1-(0.44^{2} + 0.56^{2}) = 0.49, qué está cerca del máximo el posible índice de diversidad de 0.5 (o generalmente 1/n donde "n" es el número de categorías). El valor límite, 0.5 se alcanza cuando cada clase tiene exactamente el mismo número de miembros y por consiguiente exactamente la misma probabilidad de ser escogió.

La fórmula del índice de diversidad para un target binario es 2p1 (1-p1) donde p1 es la probabilidad de la primer clase. Usando esta fórmula, después de la primer división, el primer nivel tiene 5 cuadrados y 1 triángulo, dando un valor de diversidad de 0.28. El segundo nivel tiene 4 cuadrados y 6 triángulo con un índice de diversidad del 0.48. El promedio de las diversidades entre ambos niveles es ((0.28*6) + (0.48*10))/16 = 0.41.

La diversidad total del árbol es igual a la diversidad del nodo raíz menos el promedio de las diversidades de los niveles. En este caso, la división X>50 ha reducido la diversidad total desde 0.49 a 0.41.

La medida de diversidad es calculada para dos particiones, y el mejor divisor o predictor es el que posee la disminución más grande en diversidad. Esto se repite para todos los campos. El "ganador" es seleccionado como el PREDICTOR para el nodo.

Expansión y crecimiento del árbol

La división inicial produce dos nodos, cada uno de los cuales es entonces dividido de la misma manera como el nodo de la raíz. Primero, si todos los resultados en el nodo son iguales, no hay ningún sentido continuar intentando la división. En este caso, el nodo se etiqueta como un nodo final.

Por otra parte, el algoritmo del árbol de decisión examina todos los campos de entrada para encontrar los posible candidatos a ser Predictores. Si el campo toma un solo valor, es eliminado ya que se estima que a partir de aquí no hay forma de crear una división. Un campo categórico que se ha usado como el primer y más alto predictor en el árbol es probablemente identificado rápidamente. Luego se van determinado los mejores divisores para cada uno de los campos y cuando no puede encontrarse ningún predictor que disminuya significativamente la diversidad de un nodo dado, se ha llegado entonces al nodo final.

Finalmente, cuando el árbol ha llegado a su punto máximo de expansión, describe las reglas de clasificación para el conjunto de variables que se han sido proporcionadas como input y sólo para ellas.

Recortando el Árbol –Pruning the tree-

Este proceso implica remover nodos y caminos para mejorar la performance del árbol. Un recorte del árbol es, en efecto, un subconjunto del árbol de decisión.

El árbol de decisión se mantiene en crecimiento siempre que puedan encontrarse nuevos predictores que mejoren la habilidad del árbol para separar los registros en las distintas clases. Si los datos utilizados se usaran para la evaluación, cualquier recorte del árbol sólo aumentaría la proporción del error, de la misma manera que si se quisiera aplicar el árbol para clasificar nuevos set de datos.
Los algoritmos que trabajan sobre la construcción del árbol despliegan modelos que sólo pueden ser aplicados al set de datos de origen en tanto que su utilización para la predicción puede resultar francamente perjudicial.

Hay varios acercamientos a este problema. Los árboles pueden ser árboles más pequeños, mediante la aplicación de las llamadas técnicas de bonsai, que intentan impedir el crecimiento del crecimiento del árbol más allá de lo deseado. La técnica consiste en ir aplicando varias pruebas a cada nodo para determinar si es probable trabajar con subdivisiones. La prueba puede contener simples requerimientos como números mínimo de registros que deben estar contenidos en un nodo ó puede ser más complicado involucrando pruebas estadística para testear la significatividad de la división propuesta.

La experiencia muestra, sin embargo, que poniendo un tamaño del nodo mínimo lo suficientemente grande la técnica tiende a producir árboles bien desarrollado. Esto es porque cuando los nodos contienen muchos archivos, ellos son probablemente una muestra representativa del universo

Los métodos de recorte, por otro lado, permiten el árbol de decisión inicial para crecer bastante profundo y luego cortar las ramas que no se generalizan. Actualmente, muchas técnicas de recorte comparten el inconveniente de sólo hacer uso de la información del set inicial. Una forma de aproximación es encontrar la proporción de error de clasificación asociada con varios sub-árboles más pequeño y más pequeño del árbol inicial. Por supuesto, cuando éstos ratios de error son calculados obre los mismos datos usados para la creación del árbol, el error mantiene la complejidad original del árbol y agrega la complejidad del error haciéndolo aún más complejo.. En este esquema de recorte, una rama se mantiene sólo cuando mejora ampliamente la clasificación a pesar de la complejidad extra que implica.

Estos acercamientos para recortes de los árboles son apropiados en el ambiente académico donde la abundancia de datos es limitada pero cuando los datos son abundantes como en el caso del ambiente de negocios, la mejor aproximación está basada en un recorte del árbol sobre su actual performance. La performance del árbol y todos de su sub-árboles es medida sobre un set separado de datos preclasificados, llamado set de testeo. Con un simple set de testeo, el algoritmo puede minimizar el error sobre el set de testeo. Con un set de testeo, es posible dirigirse más directamente al problema del modelo general seleccionando el sub-árbol que performa más consistentemente alrededor de varios set de pruebas.

Consecuencias de seleccionar árboles de decisión

A esta altura, habiendo evaluado como se pueden construir y modelas los árboles de decisión, conviene avanzar sobre las consecuencias para el Data Mining:

Todas las divisiones en un árbol se realizan sobre una variable, los árboles de decisión nunca descubren reglas que implican la relación entre variables.

Manejando las variables de entrada

Una de las principales ventajas de esta herramienta es que no es sensible a las diferencias de escala entre los datos de origen, datos fuera de rango (outliers) y distribuciones sesgadas, esto implica una carga de trabajo menor, en término de los datos y respecto de las técnicas de segmentación y redes neuronales oportunamente descriptas.

El manejo de variables categóricas puede causar problemas. Dependiendo del algoritmo utilizado, las variables categóricas pueden dividirse en cada valor asumido por la variable y pueden llevarse a un árbol muy espeso e inmanejable. Otros algoritmos encuentran maneras de agrupar clases de etiquetas en un número pequeño de grandes clases combinando clases que generan divisiones similares. El excesivo crecimiento de divisiones pone en riesgo la utilidad del árbol.

Árboles y Reglas

Estructura de una regla:

Los árboles de decisión son a menudo seleccionados por su habilidad para generar reglas entendibles. Y esto es verdad, ya que la forma de ubicar la clasificación del registro es tan simple como trazar el camino desde el nodo raíz donde aparece el dato y analizando cómo fluye a lo largo de las ramas del árbol generar la regla.

En la aplicación práctica podrá observarse la construcción de Reglas a través de Árboles de Decisión y su aplicación a ejemplos de negocio.

REDES NEURONALES – MODELOS PREDICTIVOS DE APRENDIZAJE

Las redes neuronales son herramientas que se utilizan para construir modelos predictivos no lineales que "aprenden" a través de entrenamiento y que se asimilan a los modelos de redes de neuronas biológicas.

Las **redes neuronales** son un procedimiento totalmente diferente de analizar los datos respecto a cualquiera de las técnica multivariantes.

En lugar de conceptualizar el problema como matemático, las redes neuronales utilizan el cerebro humano y su estructura para desarrollar una estrategia de procedimiento. Aunque es muy utópico pensar en la posibilidad de construir redes neuronales tan complejas como el cerebro humano, es posible utilizar sus principios básicos de unidades de procesamiento múltiple involucradas en el reconocimiento de pautas.

Las redes neuronales difieren no sólo en su estructura, sino también en proceso. Un elemento clave de una red neuronal es el **aprendizaje** (otra analogía con el cerebro humano), mediante el que los errores de los resultados (predicción o clasificación) se retroalimentan en el sistema y se ajustan consecuentemente. A continuación el proceso se repite, aprendiendo de cada conjunto de errores de resultado. Se trata de un proceso secuencial, una vez por instante, en comparación con otra técnica multivariante, que considera el conjunto completo de casos simultáneamente.

Las redes neuronales se hicieron operativas por primera vez en los últimos años de la década de los 50. Sin embargo, a finales de los 60, la investigación demostró que las redes neuronales de aquel momento eran bastante limitadas en capacidad, y el proceso de investigación sufrió un estancamiento. El interés volvió en los 80 a medida que al desarrollo teórico se unió un mayor poder de cálculo. La principal mejora fueron las capas ocultas (se expondrán más detalles en una sección ulterior), que permitieron a las redes neuronales representar sistemas mucho más complejos. Hoy en día, las redes neuronales se utilizan en casi todas las disciplinas o áreas de análisis. La naturaleza flexible de las especificaciones del sistema las hace adaptables a un amplio rango de problemas, que van desde la predicción a la clasificación o incluso el análisis de series temporales.

En primer lugar, se clasificará las relaciones entre las redes neuronales y las técnicas estadísticas multivariantes. Las redes neuronales pueden tratar muchos de los problemas que tratan las técnicas multivariantes de la regresión múltiple, el análisis discriminante y el análisis de segmentos. En muchos casos, las redes neuronales producen resultados comparables, y por tanto depende del analista de datos seleccionar entre los métodos de acuerdo a los objetivos de la investigación. Asimismo, las redes neuronales tienen una base estadística subyacente, por ejemplo, el impacto de las distribuciones de *inputs* (no normales) sobre la estimación de las ponderaciones. La principal diferencia de las técnicas multivariantes es la ausencia de contrastes de inferencia estadística para las ponderaciones del ajuste global del modelo. Sin embargo, el analista de datos no debe tomar las redes neuronales como algo menos riguroso, sino más bien como una variación del enfoque.

Conceptos básicos de las redes neuronales

Las redes neuronales tienen una estructura simple y operativa que puede describirse mediante cuatro conceptos:

- El **Tipo de modelo de red neuronal**
- Las **Unidades de Procesamiento-Nodos** que recogen información, procesan dicha información y crean valor resultante
- El **Sistema de Nodos-Red** organizado para transmitir señales desde los nodos de entrada a los nodos de salida, con algunos nodos intermedios entre ellos.
- La **Función de Aprendizaje** mediante la cual el sistema "retroalimenta" los errores de predicción para recalibrar el modelo.

Se discutirán cada uno de estos conceptos y su interrelación en detalle.

Tipos de modelos de redes neuronales

Existen tres tipos de redes neuronales:

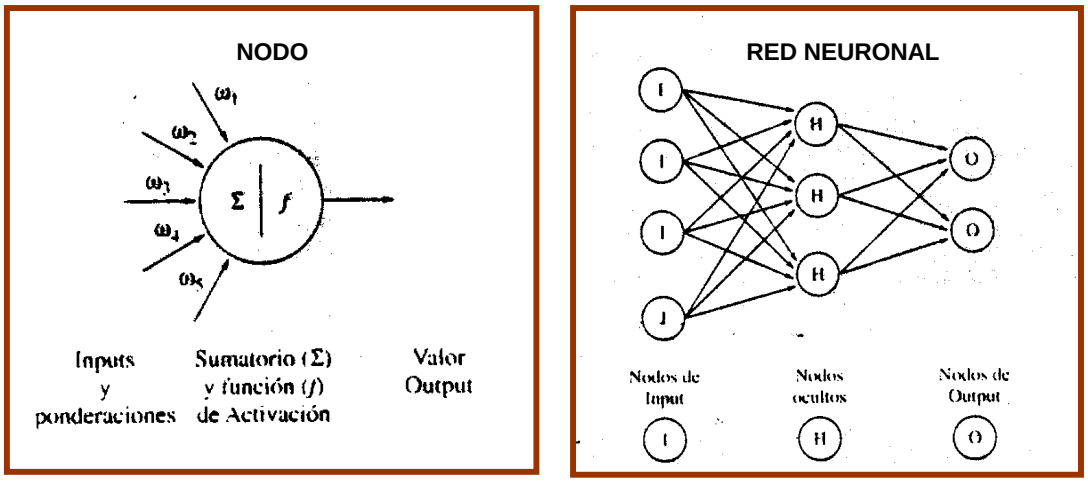
1. **El perceptor multicapa**
2. **La función de base radial**

3. Las redes de Kohonen

1. El **Modelo del perceptor multicapa** es el más común y es el tipo que se utiliza en el ejemplo que se expondrá a continuación.
2. La **función de base radial** es un método desarrollado más recientemente que se puede utilizar para las mismas tareas que el modelo multicapa pero su forma de operar es sensiblemente diferente.
3. El **modelo de Kohonen** es apropiado sólo para los problemas de segmentos.

Unidades de Procesamiento-Nodos

El elemento más básico de una red neuronal es un **nodo,** una unidad de procesamiento autocontenido que actúa en paralelo con otros nodos de una red neuronal. El **nodo** es análogo a la neurona del cerebro humano, que acepta *inputs* y genera *outputs*. Una representación simple del nodo y su forma de actuar se muestra en la figura adjunta.



Los nodos aceptan un número de *inputs* de otras fuentes (nodos). Cada conexión de otro nodo tiene un valor asignado. La primera tarea del nodo es procesar los datos de entrada creando un valor resumen en el que cada valor de entrada se multiplique por su respectiva ponderación (nótese que es la misma operación desarrollada por los método multivariantes cuando se calcula el valor teórico). Este valor resumen, se proceso a continuación mediante una **función de activación** para generar un valor de salida, que se envía al siguiente nodo del sistema. Las funciones de activación son generalmente funciones no lineales, como la **función sigmoide**, que es una clase general de curvas de forma S que incluyen la función logística.

Sistemas de Nodos-Red Neuronal

La red neuronal es una ordenación secuencial de tres tipos básicos de nodos o capas:

- ***Nodos de entrada***
- ***Nodos de salida***
- ***Nodos intermedios (escondidos)***

Los ***nodos de entrada*** reciben valores iniciales de los datos de cada caso y los transmiten a la red neuronal. Un nodo de entrada representa una única variable o pauta. Las variables métricas sólo requieren un nodo por variable. Las variables no métricas tienen que estar **codificadas**, lo que significa que cada categoría se representa por una variable binaria. Por tanto, se pueden representar tres variables no métricas mediante tres nodos de entrada binarios. La primera categoría tendría valores de 1,0,0, para las tres variables, la segunda categoría tendría valores de 0,1,0 y la tercer categoría tendría valores de 0,0,1. Un ***nodo de salida*** recibe entradas y calcula un valor de salida, pero en lugar de ir a otro nodo, éste es el valor final. Si se trata de un ***modelo predictivo***, entonces éste es el **valor predictivo**. Si el modelo se utiliza para ***clasificación***, entonces ése es el valor utilizado en los procesos de clasificación.

En casi todas las redes neuronales, existe un tercer tipo de nodo contenido en la **capa oculta**. Es un conjunto de nodos utilizados por la red neuronal para representar relaciones más complejas que las relaciones uno a uno de *inputs* a *output*. Es la capa oculta y la función de activación la que permite a las redes neuronales representar fácilmente las relaciones no lineales, que son muy problemáticas para las técnicas multivariantes.

Este diseño de redes neuronales permite a cada nodo actuar independientemente, pero en paralelo, con el resto de los nodos. Por ello, ofrece a la red neuronal una gran flexibilidad en los tipos de relaciones *input-output* que puede manejar.

Función de Aprendizaje

La característica de una red neuronal, que verdaderamente la aparta del resto de las técnicas multivariantes, es su capacidad de "aprender" o "corregirse a sí misma" basándose en sus propios errores.

Previamente, se discutieron los conceptos de ponderaciones en las conexiones entre nodos. Utilizando una analogía biológica, las ponderaciones representan un estado de memoria, el "mejor cálculo" de cómo hacer predicciones a partir de los resultados de los nodos. Una vez que la entrada de un caso se procesa a través del sistema, puede compararse con el valor del resultado efectivo. A esto se le denomina preparación del sistema en un modelo de **aprendizaje supervisado**. Los valores resultantes y los efectivos se comparan. Si existe alguna diferencia entre los dos valores (parecido a un valor residual), entonces se ajustaría el modelo con la esperanza de mejorarlo. La forma más común de preparación es la retropropagación. En esta aproximación, se calcula el error en el valor del output y se distribuye hacia atrás en el sistema. Como funciona por su vía a través del sistema de nodos, las ponderaciones cambian proporcionalmente, aumentando o disminuyendo dependiendo de la dirección del error. Una vez que todas las ponderaciones se han recalibrado, se introduce en la red el *input* de otro caso y se comienza de nuevo el proceso. El objetivo es procesar un gran número de casos a través de la red neuronal en la fase de preparación, de tal forma que pueda hacer las mejores predicciones para todas las pautas de entrada de datos. Existe también un modo **sin supervisar** en el que no se da ninguna retroalimentación hacia lo que es el valor correcto de *output.* Esta aproximación se utiliza sólo en los problemas de *clustering* porque no hay forma de conocer las soluciones efectivas del segmento.

ESTIMACIÓN DE UN MODELO DE RED NEURONAL

Ahora que se ha descripto las operaciones básicas y los componentes de la red neuronal, se centrará el desarrollo en cinco aspectos fundamentales implicados en la estimación efectiva de un modelo de red neuronal:

a) Preparación de los datos

b) Definición de la estructura del modelo

c) Estimación del modelo

d) Evaluación de los resultados del modelo

e) validación del modelo

a) Preparación de los datos

Las redes neuronales son como el resto de los métodos estadísticos en un sentido: el resultado depende de los datos ingresados. Las redes neuronales no tienen la capacidad de transformar la baja calidad o el mal acondicionamiento de los datos en un buen modelo. El analista de datos, por tanto, debe examinar los datos junto con cualquier otro método estadístico. Al preparar los datos para su uso en una red neuronal, el analista de datos debe considerar:

1. **El tamaño muestral**
2. **Las distribuciones, transformaciones y codificaciones de los datos**

1. El tamaño muestral

La primera tarea es crear las bases de datos para la estimación de su red neuronal. Se necesita una muestra de calibración, conocida como muestra de preparación, utilizada para estimar las ponderaciones y una muestra distinta de validación para evaluar independientemente la capacidad del modelo. La escisión de la muestra original está basada primordialmente en el tamaño de muestra exigido para las muestras de calibración y validación.

El tamaño muestral utilizado para calibrar un modelo de red neuronal puede tener tanto impacto sobre los resultados como con cualquier otra técnica multivariante. En primer lugar, un modelo puede estar perfectamente ajustado si el tamaño muestral es menor que el número de parámetros estimados. Pero a medida que el número de casos se aproxima al número de coeficientes, se produce el sobre-ajuste y el modelo se hace muy específico para la muestra, perdiendo generalidad. El número de ponderaciones en un modelo simple de red neuronal puede aumentar rápidamente. El modelo de red neuronal de la figura tiene 12 coeficientes entre los inputs y la capa oculta y otros 6 coeficientes entre la capa oculta y los nodos de resultado por un toral de 18 ponderaciones. Por tanto, el número de ponderaciones se relaciona con el número de capas y el número de nodos en cada capa. Añadiendo cada vez un nodo de entrada se añadirían tres coeficientes, mientras que añadiendo una segunda capa oculta de tres nodos se añadirían nueve coeficientes.

El modelo requiere que haya algo más que un caso adicional al número de coeficientes. La norma habitual es tener entre 10 y 30 casos en la base de datos de calibración para cada parámetro estimado. Aunque es no es más que una sugerencia, debería considerarse seriamente cuando se diseñe la red. Habría que decir que nunca se tienen realmente muchos casos (al contrario que en los test estadísticos, en los que la potencia del test aumenta notablemente y los contrastes de inferencia estadística siempre se muestran significativos). Los casos se pueden muestrear aleatoriamente si el tamaño de muestra original se juzga que es muy grande para procesarlo o resulta inabarcable para el analista de datos.

2. Las distribuciones, transformaciones y codificaciones de los datos – Examen de los datos

Incluso aunque no se utilicen los tests de inferencia estadística, se deberá seguir examinando los datos con el fin de evaluar la simetría, la no normalidad y los casos atípicos. Respecto a las variables métricas, existen técnicas para evaluar estas características de una forma fácil y llevar a cabo las transformaciones si fuera necesario. Las variables no métricas pueden resultar problemáticas, particularmente si el nodo de *output* es una variable métrica simétrica. Si existen dos categorías no es bueno tener el 98% de los casos en una categoría y el 2% en la otra categoría. El modelo de red neuronal se concentrará en el valor grande (98%) porque el error producido por la pérdida de la otra categoría es reducido. Se deberá buscar un tamaño parecido para cada categoría. Si esto no se pudiera de forma natural, deberá verse la forma de igualar las categorías a través de la muestra. El problema es que el tamaño total de la muestra está limitado por el grupo de *output* más pequeño.

Además de la simetría y la normalidad, los datos también deberían examinarse en busca de puntos perdidos o missing. Los casos atípicos deberían examinarse cuidadosamente para su eliminación, dado que pueden ver seriamente afectados por los resultados conjuntos y los procedimientos de preparación. El último aspecto se centra en la estandarización de las variables. La escala utilizada para medir las variables puede provocar la estimación de grandes coeficientes para los valores más altos de la escala. Aunque existen muchas sugerencias sobre cuándo estandarizar, este proceso no tiene un riesgo real y se sugiere que todas las variables se estandaricen.

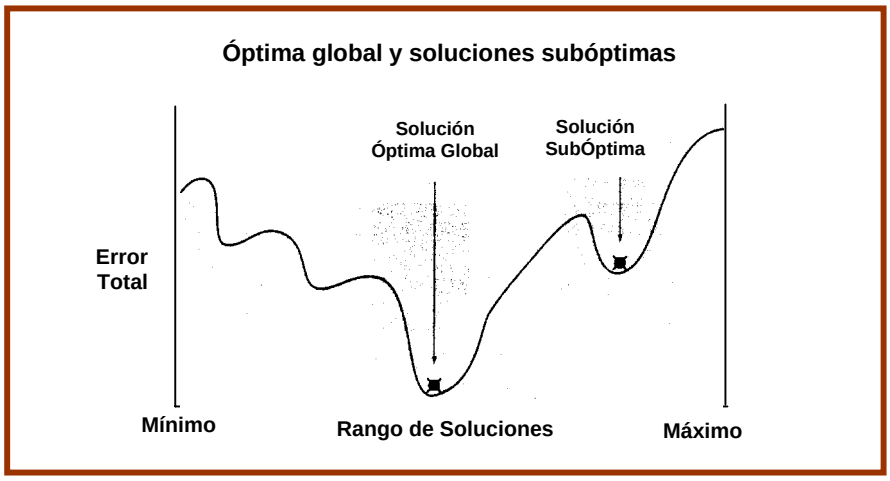
b) Definición de la estructura del modelo

Una vez que se han examinado los datos y se han realizado las transformaciones que sean pertinentes, se debe especificar la estructura del modelo. Dado que los *inputs* y los *outputs* ya se han seleccionado, en este paso se deciden el número de capas ocultas y el número de nodos en cada capa. Una capa oculta adicional puede mejorar ligeramente la estimación, pero con un riesgo mayor de sobre-ajuste y de encontrar sólo una solución sub-óptima, aumentando también el tiempo necesario para la estimación. Si se añade una segunda capa, se pueden utilizar procedimientos de validación de forma rigurosa. En lo que se refiere al número de nodos en una capa oculta, se trata de un método de ensayo y error. Sólo debería modificarse el número de nodos en la capa oculta para encontrar el mejor ajuste del modelo. Si las dos soluciones tienen igual ajuste, se seleccionará la estructura más simple (menor número de nodos). Un último elemento es el **nodo sesgado**, que es una clase de nodo oculto pero que tiene un valor constante de 1,0. Se puede añadir para que actúe como un término constante en una ecuación de regresión.

c) Estimación del modelo

El principal objetivo es conseguir el mejor ajuste posible del modelo buscando la solución global óptima y no la sobre-preparación. La solución global óptima es la mejor solución posible de todas las soluciones posibles, denominadas el **espacio de búsqueda**. Dado que el proceso de estimación es un procedimiento iterativo y que busca modos de ajuste incrementales, es posible que no siempre se encuentra la solución óptima.

La **-Figura Óptima global y soluciones sub-óptimas-** adjunta representa una solución posible sub-óptima y una solución óptima global.



Si el punto de partida estuviera más a la derecha, la solución puede mejorarse a medida que se mueve hacia la izquierda, pero cuando se alcanza el fondo del "valle" no podría hacer ninguna mejora. El proceso no tiene ninguna forma de "saltar" el valle, de tal forma que puede proceder hasta el óptimo global. Sin embargo, si el punto de partida estuviera a la izquierda, se alcanzaría el óptimo global. Esto ilustra la necesidad de utilizar puntos de partida múltiples para asegurar que ha alcanzado el óptimo global.

Sabido es que todos los datos tienen algún error o ruido aleatorio, así que no puede esperarse un ajuste perfecto. Pero si un modelo se entrena con muchos casos, es posible que se sobre-entrene, lo que significa que se han explicado todas las relaciones básicas entre los nodos y que comienza a representar el error aleatorio del modelo. Este problema se puede evitar de dos formas. La primera es fijar un límite inferior de error y detener el entrenamiento cuando se alcance. Aunque sea arbitrario, evita directamente el problema. Un segundo enfoque, es controlar la tasa de error tanto para las muestras de calibración como de validación. Cuando la red neuronal se está preparando inicialmente, el ajuste mejorará ambas muestras así como las ponderaciones se calibran. Pero en un punto, el error de la muestra de validación se nivelará o incluso empezará a aumentar divergiendo de la muestra de calibración. Éste es el punto en el cual la muestra de calibración se empieza a sobre preparar, es decir se hace muy específica para la muestra (y no generalizable). Esto también produce un bajo ajuste de la muestra de validación. Cuando se haya encontrado la mejor de las soluciones posibles la investigación debería detenerse.

d) Evaluación de los resultados del modelo

La evaluación de un modelo de red neuronal consiste fundamentalmente en la evaluación del nivel de predicción o clasificación de las variables de entrada. Por ejemplo, en un problema de clasificación, la matriz de clasificación tabula de forma cruzada los valores actuales y previstos situando las predicciones corregidas en la diagonal. Desde este número, se puede calcular el porcentaje correctamente clasificado y comparar con diferentes criterios. Si el *output* es una variable métrica, entonces la medida común de ajuste es el error cuadrático medio.

Incluso aunque un modelo de red neuronal pueda desarrollar tareas comparables a las otras técnicas multivariantes, tales como regresión múltiple o el análisis discriminante, no ofrece información sobre la importancia relativa de las variables de entrada. Aunque existen ponderaciones para cada variable de entrada, no son directamente interpretables porque también deben combinarse con las ponderaciones de la capa oculta de la variable de salida. Hasta el momento no hay un método simple de interpretación de la solución en relación con la importancia del impacto de una única variable de entrada.

e) Validación del modelo

El paso final es validar la solución para asegurar que es el óptimo global y que es tan generalizable como sea posible. Como ya se ha mencionado, es esencial que se cree una muestra de validación para ofrecer una evaluación independiente del ajuste del modelo que no dependa de la muestra de calibración. Mientras sea posible, se debería emplear una nueva muestra de casos para una evaluación adicional del ajuste. La estabilidad de la solución se evalúa ofreciendo diferentes puntos de partida para las ponderaciones y reorganizando el orden de los casos de la muestra de calibración. Finalmente, también se debería variar el número de nodos para asegurarse de que no es posible una mejor solución.

En síntesis:

El proceso de estimación de un modelo de red neuronal no se lleva a cabo en un único paso, sino que requiere de la evaluación constante de las soluciones del modelo a efectos de ganar precisión predictiva, vigilar el sobre-entrenamiento o el sobre-ajuste. Se debería "experimentar" con la solución en la medida en que sea posible porque el procedimiento no siempre garantiza una solución óptima global o general. Cualquier analista de datos que desee utilizar este procedimiento debería ser consciente de las consideraciones a cada paso, dado que no es un método en el que baste con utilizar valores por "defecto" y estar seguro de obtener un modelo aceptable.

Lecciones aprendidas:

- **Las redes neuronales son adecuadas para problemas de tipo predictivo.**
- **Un problema apropiado para una red neuronal tiene tres características:**
  1. Se comprenden claramente los INPUTS
  2. Se comprende claramente el OUTPUT
  3. Existen ejemplos (experiencia) suficientes para entrenar a la red
- **Un modelo neuronal es tan bueno como lo es el set de datos usado para entrenar la red**
- **El modelo es estático y debe ser explícitamente actualizado agregando ejemplos recientes y re-entrenando la red para asegurar su vigencia y utilidad**
- **La red neuronal no produce reglas explícitas que describan el modelo**
- **Con modelos neuronales se puede atacar una gran variedad de problemas y producir buenos resultados aún en dominios complejos con variables continuas y categóricas**
- **Son apropiados para tareas de clasificación y predicción cuando los resultados del modelo son más importantes que comprender cómo funciona el modelo.**

UTILIZACIÓN DE UNA RED NEURONAL PARA LA CLASIFICACIÓN

Una de las aplicaciones más comunes de las redes neuronales son los problemas de clasificación: decidir a qué grupo pertenece una observación. Esto se corresponde al análisis discriminante y al problema de regresión logística. Para ofrecer una comparación entre las capacidades el análisis discriminante y las redes neuronales, en este apartado se revisará el problema de los dos grupos del análisis discriminante. Además de los resultados del análisis discriminante, también se estimará cuatro modelos de red neuronal. Los tres primeros, tienen una capa oculta con cuatro, seis y ocho nodos. El cuarto modelo de red neuronal tiene dos capas ocultas con cuatro nodos cada una. Los modelos de red neuronal alternativos se estiman para averiguar la precisión predictiva que se gana a través del aumento de la complejidad del modelo.

La **Tabla RN1: Comparación de la red neuronal y los modelos de análisis discriminantes. Dos grupos de clasificación** contienen las descripciones del modelo de red neuronal y los resultados predictivos. La precisión de la clasificación se mide por el porcentaje de la clasificación correcta.

TablaRN1: Comparación de la red neuronal y los modelos de análisis discriminantes. Dos grupos de clasificación

Modelo de red neuronal				
Modelo	Estructura [a]	Número de ponderaciones	Error de preparación [b]	Precisión de clasificación [c]
NN1	7-4-1	32	0.196	80%
NN2	7-6-1	48	0.142	90%
NN3	7-8-1	64	0.026	100%
NN4	7-4-4-1	56	0.034	100%

Análisis discriminante

Modelo de dos grupos	*97.5%*

[a] El primer valor es el número de nodos de entrada. El último valor es el número de nodos de salida. Los valores de en medio representan los nodos en las capas ocultas
[b] Error cuadrático medio
[c] Porcentaje correctamente clasificado en la muestra de contraste (red neuronal) ó muestra de validación (análisis discriminante)

En primer lugar, el modelo del análisis discriminante predice muy bien, con un porcentaje del 97,5% de acierto. Pero el modelo de red neuronal supera al análisis discriminante para mayores grados de complejidad del modelo. Los niveles de precisión de clasificación predictiva de los modelos de red neuronal van desde el 80% del modelo más simple al 100% de los modelos más complejos. Nótese que aunque no era posible ganar en precisión predictiva para los dos modelos más elevados, la suma de la segunda capa oculta realmente aumenta la tasa de error de preparación. Esto apoya la tesis de que con una única capa oculta es suficiente en la mayoría de los casos.

Se debería considerar la aplicación de redes neuronales para los problemas de predicción y clasificación, especialmente cuando se hace más énfasis en la precisión de la clasificación y no en la interpretación del valor teórico (*valor teórico = combinación lineal de variables compuesta a partir de ponderaciones empíricas aplicadas a un conjunto de variables especialmente determinadas*). Como se ha visto, las redes neuronales pueden ajustarse e incluso superar la capacidad predictiva de las técnicas multivariantes apropiadas. La creciente capacidad del software, como NURAL CONCCECTION by SPSS, facilita enormemente la adopción de esta técnica.

Resumen

Las redes neuronales presentan una aproximación del análisis de datos, cuya capacidad para manejar relaciones complejas, particularmente aquellas de naturaleza no lineal, ofrecen un instrumento analítico de gran capacidad para los tipos de problemas que pueden tratarse. Esta flexibilidad proporciona la base de una superior estimación de resultados en muchos problemas de predicción y clasificación. No ofrece, sin embargo, interpretación de la importancia relativa de las variables de entrada no de sus interconexiones. Por lo tanto, se debería emplear redes neuronales en aquellas situaciones en las que:

1. ***Las técnicas multivariantes no puedan tratar relaciones complejas***
2. ***La clasificación u la predicción sean los objetivos primordiales.***

ANÀLISIS DE REGRESIÒN – MODELOS DE PREDICCIÓN

El **análisis de regresión múltiple** es una técnica estadística que puede utilizarse para analizar la relación entre una única **variable criterio** (**criterio**) y varias **variables independientes** (**predictores**).

El **objetivo del análisis de regresión múltiple es usar las variables independientes cuyos valores son conocidos para predecir la única variable criterio seleccionada**. Cada variable predictor es ponderada, de forma tal que las ponderaciones indican su contribución relativa a la predicción conjunta. Al calcular las ponderaciones, el procedimiento del análisis de regresión asegura la máxima predicción a partir del conjunto de variables independientes. Estas ponderaciones facilitan también la interpretación de la influencia de cada variable en la realización de la predicción, aunque la correlación entre las variables independientes complica el proceso de interpretación. El conjunto de variables independientes ponderadas es conocido también como valor teórico de la regresión, una combinación lineal de las variables independientes que predice mejor la variable criterio. La ecuación de regresión, también denominada como el valor teórico de la regresión, es el ejemplo de valor teórico más ampliamente reconocido entre todas las técnicas multivariantes.

El **análisis de regresión múltiple** es una técnica de **dependencia**. Por tanto, al utilizarla, se deben dividir las variables entre variables dependientes y variables independientes. El análisis de regresión es también una herramienta estadística que debería utilizarse sólo cuando tanto las variables dependientes como las independientes son métricas. Sin embargo, bajo ciertas circunstancias, es posible incluir datos no métricos para las variables independientes (transformando los datos ordinales o los nominales en variables ficticias) o la variable criterio (mediante el uso de una medida binaria en la técnica especial de la regresión logística). En resumen, al aplicar el análisis de regresión múltiple:

1. ***Los datos deben ser métricos o apropiadamente transformados.***

2. ***Antes de derivar la ecuación de regresión, se debe decidir qué variable será dependiente y cuál de las restantes variables serán independientes.***

Regresión múltiple y regresión simple

El objetivo del análisis de regresión es predecir una única variable criterio a partir del conocimiento de una o más variables independientes.

Cuando el problema implica una única variable independiente, la técnica estadística se denomina **regresión simple**.

Cuando el problema implica dos o más variables independientes, se denomina **regresión múltiple.**

Un ejemplo de regresión simple y múltiple
Para ilustrar los principios básicos relacionados con la utilización de esta técnica, se proporcionan los resultados de un pequeño estudio de ocho familias y su uso de tarjetas de crédito. Se identificaron factores potenciales (tamaño familiar, ingresos familiares y el número de automóviles poseídos) y se recogieron datos de cada una de las ocho familias, los cuales se detallan en **Tabla AR1-Perfil de Variables**:

Tabla AR1-Perfil de Variables

Id Familia	*Número de posesión de tarjeta de crédito* Y	*Tamaño de la familia* V_1	*Renta familiar ($000)* V_2	*Número de posesión de automóviles* V_3
1	4	2	14	1
2	6	2	16	2
3	6	4	14	2
4	7	4	17	1
5	8	5	18	3
6	7	5	21	2
7	8	6	17	1
8	10	6	25	2

En la terminología del análisis de regresión, la variable criterio (Y) es el Número de tarjetas de crédito utilizado y las tres variables (V_1, V_2 y V_3) representan el tamaño de la familia, los ingresos familiares y el número de automóviles poseídos, respectivamente. La exposición del ejemplo se divide en tres partes para entender de qué manera la regresión estima la relación entre la variable independiente y la variable criterio.

Los temas a tratar son:

1. **Predicción sin una variable independiente, utilizando sólo una medida única –la media.**
2. **Predicción utilizando una única variable independiente –regresión simple.**
3. **Predicción utilizando varias variables independientes –regresión múltiple.**

1. Predicción sin variable independiente

Antes de hacer la estimación con la primera ecuación de regresión, se empezará con el cálculo de una línea básica con la que comparar la capacidad de predicción de los modelos de regresión. La línea básica debe representar la mejor predicción sin el uso de variables independientes. Podría utilizarse cualquier número de opciones (por ejemplo: la predicción perfecta, un valor especificado previamente ó una de las medidas de tendencia central, como la media, la mediana o la moda), sin embargo la línea predictor utilizada en la regresión es la media simple de la variable dependiente, lo cual tienen varias propiedades deseables. En el ejemplo, la media aritmética del número de tarjetas utilizadas es siete. La predicción sería "*el número medio de tarjetas de crédito mantenidas por una familia es siete*". En forma de ecuación de regresión se diría:

Predicción del número de tarjetas de crédito = Número medio de tarjetas de crédito

$$\hat{Y} = \bar{y}$$

Dado que la media no dará una predicción perfecta de cada valor de la variable criterio, se tiene que crear alguna manera de valorar al exactitud de predicción, que se podría usar tanto con la predicción de la línea de base como con los modelos de regresión que vayan a crear. El modo habitual de evaluar la adecuación de una variable predictor es examinar los errores en la predicción de la variable criterio cuando se usa para la predicción. En el ejemplo, la predicción es que cada familia usa siete tarjetas de crédito, de forma tal que se está sobreestimando el número de tarjetas de crédito para la Familia 1 en tres. –Véase **Tabla AR2-Modelo de Predicción**-. Por tanto, el error es +3.

Valor teórico de regresión: Y=y
Ecuación de predicción: Y=7

Tabla AR2-Modelo de Predicción

Id Familia	*Número de posesión de tarjeta de crédito* Y	*Predicción de la línea base*[a]	*Error de la predicción*[b]	*Error de la predicción elevado al cuadrado*
1	4	7	-3	9
2	6	7	-1	1
3	6	7	-1	1
4	7	7	0	0
5	8	7	+1	1
6	7	7	0	0
7	8	7	+1	1
8	10	7	+3	9
Total	56		0	22

[a] Número medio de tarjetas de crédito utilizadas → 56 / 8 = 7
[b] Error de predicción referido al valor real de la variable dependiente menos el valor de predicción

Si este procedimiento fuese seguido para cada familia, algunas estimaciones serían demasiado altas, otras demasiado bajas y a la vez otras podrían ser correctas. Aunque se podría esperar la obtención de una medida útil de exactitud de predicción con una simple suma de los errores, esto no sería de utilidad porque el resultado sería siempre cero. Por lo tanto, la suma simple de errores nunca cambiaría, independientemente del grado de éxito que se tuvo con la predicción de la variable criterio con el uso de la media. Para solucionar este problema, se eleva al cuadrado el error y se suman los resultados. El total, denominado como la suma de los errores al cuadrado (SSE) proporciona una medida de precisión predictiva que varía según la cantidad de errores de predicción. Cuanto más pequeño es el resultado más precisas serán las predicciones.

Se eligió la media aritmética porque siempre producirá una suma de errores al cuadrado más pequeña que cualquier otra medida de tendencia central, incluida la mediana, la moda, cualquier otro valor único o cualquier otra medida estadística más sofisticada. Por lo tanto, para la encuesta de ocho familias, la utilización de la media como línea básica de predicción proporciona el mejor predictor único del número de tarjetas de crédito con una suma de errores al cuadrado de 22. En la presentación de la regresión simple o múltiple se utilizará esta predicción partir de la media como argumento para la comparación, dado que representa la mejor predicción sin utilizar variables independientes.

2. Predicción utilizando una única variable independiente –regresión simple

La regresión simple es otro procedimiento para predecir datos (al igual que la media) y utiliza la misma regla –minimizar la suma de errores cuadrados de la predicción. Se sabe que sin utilizar el tamaño de las familias se puede describir mejor el número de tarjetas de crédito mantenidas como el valor de la medio igual a siete. El objetivo para la regresión simple es encontrar una variable independiente que mejore la predicción de la línea base.

El papel del coeficiente de correlación

Utilizando la información sobre el tamaño de las familias, se podría intentar mejorar las predicciones reduciendo los errores de predicción. Para hacerlo, los errores de predicción en el número de tarjetas de crédito mantenidas ***debe*** estar asociado (correlacionado) con el tamaño de la familia. El concepto de correlación, representado por el **coeficiente de correlación (r)**, es fundamental para el análisis de regresión y describe la relación entre dos variables. Se dice que dos variables están correlacionadas si los cambios en una variable están asociados con los cambios en la otra variable. De esta forma, a medida que una variable cambia, se sabría cómo está cambiando la otra. Si el tamaño de la familia está correlacionado con el uso de tarjetas de crédito, se escribiría entonces la siguiente relación:

Número previsto de Tarjetas de crédito	**=**	**Cambio en el número de Tarjetas mantenidas asociadas Con cambio unitario en V_1**	*****	**Valor de V_1**

$$\hat{Y} = b_1 V_1$$

En la siguiente **-Tabla AR3-** se muestra una ilustración del procedimiento para algunos datos hipotéticos con una única variables independiente X_1. Si se encuentra que conforme aumenta X_1 en una unidad, aumenta la variable criterio (sobre la media) por dos, entonces se podrían hacer predicciones para cada valor de la variable independiente. El valor de predicción siempre es dos veces el valor de X_1 ($2X_1$). Sin embargo se encuentra que la predicción está mejorada por la adición de un valor constante.

En la **Parte a) de la Tabla** puede observarse que la simple predicción de dos veces X1 es errónea en cada caso. Por tanto si se cambia la descripción para añadir una constante de dos a cada predicción, proporciona predicciones perfectas en todos los casos – **Parte b) de la Tabla**. Se observará que cuando se estima una ecuación de regresión, normalmente merece la pena incluir una constante.

Tabla AR3- Mejorando la exactitud de predicción con la adición de una constante en una ecuación de regresión

Tabla AR3 - PARTE a)-PREDICCIÓN SIN LA CONSTANTE
Ecuación de predicción: $Y=2X_1$

Valor de X_1	*Variable dependiente*	*Predicción*	*Error de predicción*[b]
1	4	2	2
2	6	4	2
3	8	6	2
4	10	8	2
5	12	10	2

Tabla AR3 - PARTE b)-PREDICCIÓN CON UNA CONSTANTE DE 2.0
Ecuación de predicción: $Y=2.0+2X_1$

Valor de X_1	*Variable dependiente*	*Predicción*	*Error de predicción*[b]
1	4	4	0
2	6	6	0
3	8	6	0
4	10	8	0
5	12	10	0

Especificación de la ecuación de regresión simple

Se puede seleccionar la "mejor" variable independiente en este estudio en base a los coeficientes de correlación dado que cuanto más alto es el coeficiente de correlación, más fuerte es la relación y por lo tanto más grande es la exactitud de la predicción. La **TABLA AR 4: MATRIZ DE CORRELACIÓN POR EL ESTUDIO DE USO DE TARJETAS DE CREDITO** contiene una matriz de correlaciones entre la variable criterio (Y) y las variables independientes (V_1, V_2 o V_3).

Tabla AR4: Matriz de correlación por el estudio de uso de Tarjetas de Crédito

Variable	**Y**	**V_1**	**V_2**	**V_3**
Y-Número de tarjetas de crédito utilizadas	1.000			
V_1- Tamaño de familia	0.866	1.000		
V_2 – Renta familiar	0.829	0.673	1.000	
V_3 – Número de automóviles	0.342	0.192	0.301	1.000

Observando la primer columna, puede verse que V1 el tamaño de familia, tiene la correlación más alta con la variable criterio y por lo tanto es la mejor candidata para la primera regresión simple. La matriz de correlación también contiene las correlaciones entre las variables independientes, aspecto muy importante en la regresión múltiple.
Ahora se puede estimar el primer modelo de regresión simple para la muestra de ocho familias y ver cómo se ajusta la descripción a los datos.

Número previsto de Tarjetas de crédito mantenidas	**=**	**Constante**	**+**	**Cambio en el número de Tarjetas de crédito con diferentes tamaños de familia**	*****	**Tamaño de familia**

$$\hat{Y} = b_0 + b_1 V_1$$

En la ecuación de regresión, se representa la constante como b_0 y la denominación de b_1 se llama **coeficiente de regresión**, denotando el cambio estimado en la variable criterio por un cambio unitario de la variable independiente. El **error de predicción**, la diferencia entre los valores reales y de predicción de la variable criterio se denomina **residuo** (e). El análisis de regresión también permite que las pruebas estadísticas de las constantes y los coeficientes de regresión puedan determinar si son sustancialmente diferentes de cero (es decir, que tienen un impacto diferente de cero).

Utilizando el procedimiento matemático de mínimos cuadrados, se pueden estimar los valores de b_0 y b_1 de tal forma que la suma de los errores cuadrados de la predicción se minimiza. Para este ejemplo, los valores apropiados son una constante (b_0) de 2.87 y un coeficiente de regresión (b_1) de 0.97 por tamaño de familia. La ecuación indica que por cada miembro adicional de familia, la posesión de tarjetas de crédito es más alta como media un 0.97. Sólo se puede interpretar la constante 2.87 dentro de la gama de valores para la variable independiente. En este caso, un tamaño de familia de cero no es posible, por lo que la constante por sí sola no tiene un sentido práctico. Sin embargo, esto no anula su uso, dado que ayuda en la predicción de uso de tarjetas de crédito para cada tamaño de familia posible (en el ejemplo de 1 a 5). En los casos en los que las variables independientes pueden adquirir valores de cero, la constante tiene una interpretación directa. Para algunas situaciones especiales donde se conoce que la relación específica pasa por el origen, la denominación de constante podría ser eliminada (denominado "regresión en el origen"). En estos casos, la interpretación de los residuos y los coeficientes de regresión cambia ligeramente. En **Tabla AR 5** se muestra la ecuación de regresión simple y las predicciones y residuos para cada una de las ocho familias.

Dado que se ha utilizado el mismo criterio (minimizar la suma de los errores al cuadrado ó **mínimos cuadrados**), puede determinarse si el conocimiento del tamaño familiar ha ayudado a predecir mejor la posesión de tarjetas de crédito cuando se compara la predicción de regresión simple con la predicción de la línea básica. La suma de los errores al cuadrado utilizando la media era 22. Ahora, la suma de los errores al cuadrado es 5.5. Utilizando el procedimiento de los mínimos cuadrados y una única variable independiente, se obtiene una nueva aproximación, la regresión simple, que resulta mejor que usar sólo la media.

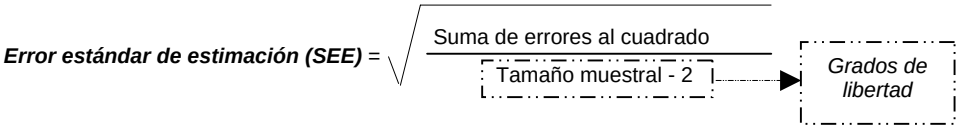
Tabla AR 5
Valor teórico de regresión: $Y=b_0 + b_1V_1$
Ecuación de predicción: $Y=2.87 + 0.97V_1$

Id Familia	*Número de posesión de tarjeta de crédito utilizadas*	*Tamaño de familia V_1*	*Predicción de regresión simple*	*Error de predicción*	*Error de predicción al cuadrado*
1	4	2	4.81	-0.81	0.66
2	6	2	4.81	1.19	1.42
3	6	4	6.75	-0.75	0.56
4	7	4	6.75	0.25	0.06
5	8	5	7.72	0.28	0.08
6	7	5	7.72	-0.72	0.52
7	8	6	8.69	-0.69	0.48
8	10	6	8.69	1.31	1.72
Total	56				**5.50**

La creación de un intervalo de confianza para la predicción

Dado que no se puede conseguir predicciones perfectas de la variable dependiente, podríase desear estimar el rango de valores que la variable a predecir puede tomar, en lugar de basarse exclusivamente en una estimación simple (puntual). La estimación puntual es la mejor estimación de la variable dependiente y puede demostrarse que va a ser la mejor predicción para cualquier valor dado de la variable independiente. Utilizando esta estimación puntual, puede calcularse el rango de valores a predecir basándose en una medida de los errores de predicción que se espera realizar. Conocido como el **error estándar de la predicción (SEE)**, esta medida es sencillamente la desviación estándar de los errores de predicción.

Utilizando la estimación puntual, puede añadirse un cierto número de errores estándar de la estimación (dependiendo del nivel de confianza deseado y del tamaño muestral) para establecer los límites superiores e inferiores de las predicciones hechas con cualquier variable (s) independiente (s). El error estándar de la estimación (SEE) se calcula mediante:



El número de SEE utilizados para derivar el intervalo de confianza se determina por el nivel de significación (alfa) y el tamaño muestral N, que da un valor t. En el ejemplo de la regresión simple, SEE=0.957 (raíz cuadrada de 5.50 dividido (8-2)). Se construye el intervalo de confianza para las predicciones seleccionando el número de errores estándar a añadir (mas / menos) mediante la búsqueda en una tabla de distribución t y la selección del valor para un nivel de confianza dado con 6 grados de libertad es 2.447. La cantidad añadida (mas / menos) al valor previsto es entonces (0.957 * 2.447) ó 2.34. Sustituyendo el tamaño medio de las familias (4.25) en la ecuación de regresión, entonces el valor previsto es 6.99 (difiere en una centésima de la media = 7). El rango esperado va entonces de 4.65 (6.99-2.34) a 9.33 (6.99+2.34)

Valoración de la exactitud de predicción

Si la **suma de los errores al cuadrado (SSE)** representa una medida de los errores de predicción, debería determinarse también una medida del éxito predictivo, que será llamada **suma de los cuadrados de la regresión (SSR)**. Conjuntamente estas dos medidas deberían igualar a **la suma total de los cuadrados (TSS),** el mismo valor que la predicción de la línea base. En la medida que se añaden variables independientes, el **total de la suma de los cuadrados** puede dividirse en (**1**) **la suma de los cuadrados prevista por la variable independiente**, también conocida como la **suma de los cuadrados de la regresión** y (**2**) la **suma de los errores al cuadrado**.

Puede utilizarse esta división para aproximar hasta qué punto el valor teórico de la regresión describe la posesión familiar de tarjetas de crédito. La predicción de la línea básica es 22; la predicción bajo el modelo de regresión con una variable independiente es 5.50. Puede entonces observarse el alcance del modelo con la investigación de esta mejora.

Suma de los errores al cuadrado (predicción de la línea básica)	***(SST)***	***-> 22.0***
menos		
Suma de los errores al cuadrado (regresión simple)	***(SSE)***	***-> 5.5***
Suma de los errores al cuadrado explicados (regresión simple)	***(SSR)***	***-> 16.5***

Por lo tanto, se explica 16.5 errores al cuadrado cambiando de la media a un modelo de regresión utilizando el tamaño de la familia. Esto supone una mejora del 75% (16.5/22=0.75) sobre el uso de la línea básica. Otra forma de expresar este nivel de precisión predictiva es el coeficiente de determinación (R^2), el ratio de la suma al cuadrado de la regresión sobre el total de la suma de los cuadrados como muestra la siguiente ecuación:

$$\textit{Coeficiente de determinación } (R^2) = \frac{\text{Suma de los cuadrados de la regresión}}{\text{Suma del total de los cuadrados}}$$

Teniendo en cuenta que el modelo de regresión que utiliza el tamaño de la familia predice perfectamente todas las tarjetas de crédito mantenidas por las familias, R^2=1.0 y que la utilización del tamaño de la familia no ofreció mejores predicciones que utilizando la media, R^2=0, cuando la ecuación de la regresión contiene más de una variable independiente, el valor de R^2 representa el efecto combinado del valor teórico en el conjunto de la predicción. El valor del R^2 es simplemente la correlación al cuadrado de los valores reales y los valores previstos.

Cuando se utiliza el coeficiente de correlación (r) para evaluar la relación entre las variables dependientes e independientes, el signo del coeficiente de correlación (+r, -r) denota la pendiente de la línea de regresión. Sin embargo, la "fuerza" de la relación se representa mejor por el R^2, que es por supuesto, siempre positiva. En el ejemplo, R^2= 0.75, que indica que el 75% de la variación en la variable dependiente se explica por la variable independiente. Cuando se menciona la variación de la variable dependiente, se refiere a la suma total de cuadrados que el análisis de regresión intenta predecir con una o más variables independientes.

3. Predicción utilizando varias variables independientes: Análisis de regresión múltiple

Se ha demostrado cómo una regresión simple mejora la predicción de tarjetas de crédito. Usando los datos del tamaño de familia, se predijo el número de tarjetas de crédito que poseería una familia, mejor

de lo que se podría realizar utilizando simplemente la media aritmética. Este resultado pone de manifiesto la cuestión de si se pudiera mejorar la predicción utilizando datos adicionales obtenidos de las mismas familias ¿mejoraría la predicción del número de tarjetas de crédito si se utilizaran datos del tamaño de la familia y además datos de otra variable, quizá la renta familiar?

El impacto de la multicolinealidad

La capacidad de una variable independiente adicional de mejorar la predicción de una variable criterio tiene relación no sólo con la correlación con la variable dependiente, sino también respecto de las correlaciones de las variables independientes adicionales en función de las variables independientes ya presentes en la ecuación de regresión. ***Colinealidad*** es la asociación, medida como correlación, entre dos variables independientes. ***Multicolinealidad*** se refiere a la correlación entre tres o más variables independientes (evidenciada cuando se hace la regresión de una respecto de las otras).

El impacto de la multicolinealidad consiste en reducir el poder predictivo de cualquier variable independiente individual en la medida en que está asociado con las otras variables independientes. Conforme se incrementa la colinealidad, la varianza única explicada por cada variable independiente se reduce y el porcentaje de predicción compartida aumenta. Dado que sólo se puede calcular una vez que se ha realizado la predicción compartida, la predicción global aumenta mucho más lentamente conforme se añaden variables independientes con un nivel de multicolinealidad alta.

La ecuación del análisis de regresión múltiple

Para mejorar aún más la predicción de la posesión de tarjetas de crédito, se utilizarán datos adicionales obtenidos de las muestras. La segunda variable independiente para ser incluida en el modelo de regresión es la renta familiar (V_2) que tiene la siguiente correlación más alta con la variable dependiente. Aunque V_2 tiene bastante correlación con V_1, es todavía la siguiente mejor variable para entrar porque V_3 tiene una correlación mucho más baja con la variable dependiente

$$\textit{Predicción del número de tarjetas de crédito utilizadas} = b_0 + b_1V_1 + b_2V_2 + e$$

donde:
$\mathbf{b_0}$ = número constante de tarjetas de crédito independientemente del tamaño y renta familiar
$\mathbf{b_1}$ = cambio en la posesión de tarjetas de crédito asociado a un cambio unitario en el tamaño familiar
$\mathbf{b_2}$ = cambio en la posesión de tarjetas de crédito asociado a un cambio unitario en la renta familiar
$\mathbf{V_1}$ = tamaño de la familia
$\mathbf{V_2}$ = renta de la familia

El modelo de regresión múltiple con dos variables independientes, cuando se estima con el procedimiento de mínimos cuadrados, tiene una constante de 0.482 con unos coeficientes de regresión de 0.63 y 0.216 para V1 y V2 respectivamente. De nuevo puede hallarse el residuo en la predicción de Y, restando de la predicción el valor efectivo. Elevando al cuadrado el error de predicción resultante se obtiene los valores expresados en la **Tabla AR 6.**

Tabla AR 6
Valor teórico de regresión: $Y=b_0 + b_1V_1 + b_2V_2$
Ecuación de predicción: $Y= 0.482 + 0.63V_1 + 0.216V_2$

Id familia	*Número de tarjetas de crédito utilizadas*	*Tamaño de familia V_1*	*Renta familiar V_2*	*Predicción de regresión simple*	*Error de predicción*	*Error de predicción al cuadrado*
1	4	2	14	4.76	-0.76	0.58
2	6	2	16	5.20	0.80	0.64
3	6	4	14	6.03	-0.03	0.00
4	7	4	17	6.68	0.32	0.10
5	8	5	18	7.53	0.47	0.22
6	7	5	21	8.18	-1.18	1.39
7	8	6	17	7.95	0.05	0.00
8	10	6	25	9.67	0.33	0.11
Total						**3.04**

La suma de los errores al cuadrado es de 3.04 para nuestra predicción utilizando tanto la renta familiar como el tamaño de la familia. Se puede comparar con el valor del modelo de regresión simple de 5.50 utilizando sólo el tamaño familiar para la predicción.
Cuando se añade la segunda variable, el R^2 aumenta a 0.86:

$$\mathbf{R^2_{(tamaño\ familiar\ +\ renta\ familiar)}} = \frac{22.0\text{-}3.04}{22.0} = \frac{18.96}{22.0} = \mathbf{0.86}$$

Esto significa que la inclusión de la renta familiar en el análisis de regresión aumenta la predicción en un 11% (0.86-0.75) debido al incremento de la potencia predictiva de la renta familiar.

La adición de una tercera variable independiente

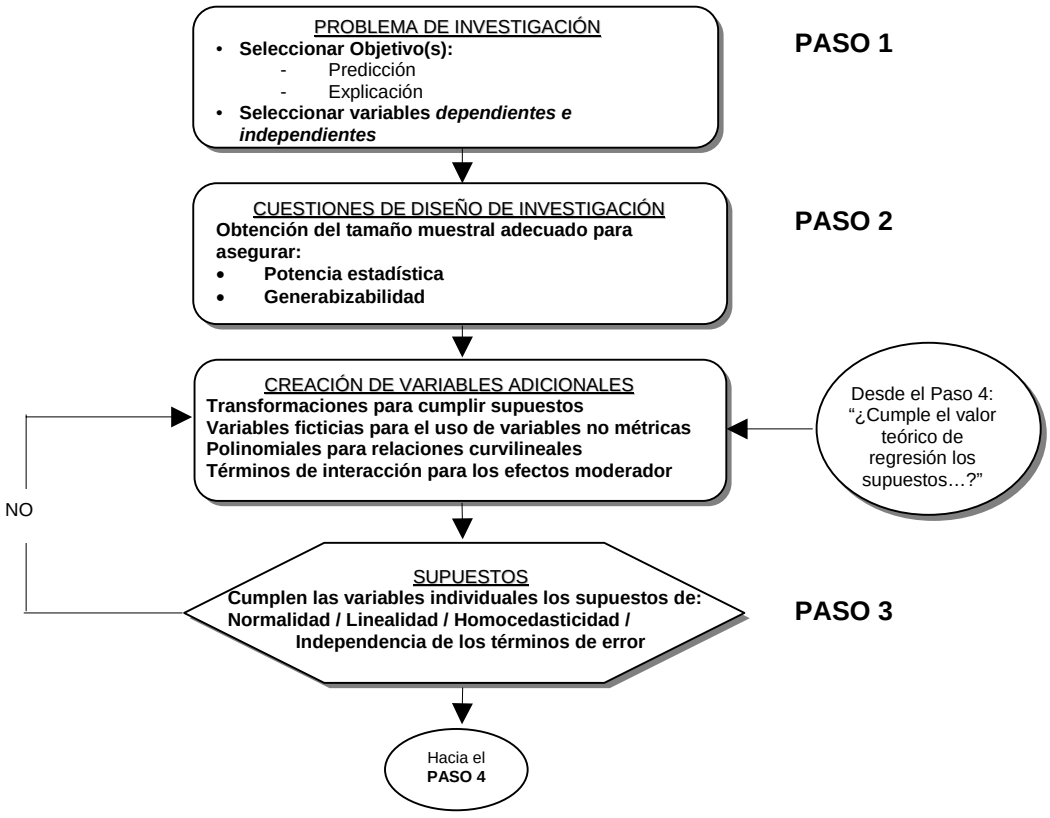
Se ha observado un incremento en la exactitud de predicción que se gana con el cambio de la ecuación de regresión simple a la ecuación de regresión múltiple, pero también se tiene que tener en cuenta que en algún momento la adición de variables independientes también será menos ventajosa e incluso en algunos casos puede resultar hasta contraproducente. En la encuesta se tiene otra adición posible de la ecuación de regresión múltiple, el número de posesión de automóviles (V_3). Si se especifica la ecuación de regresión para incluir las tres variables independientes, se puede observar una mejora en la regresión pero de menor envergadura. El R_2 aumenta a 0.87, lo que representa un incremento del 0.01 sobre el modelo anterior de regresión múltiple. Además, el coeficiente de regresión para el V_3 no es estadísticamente significativo. Por lo tanto será mejor emplear un modelo de regresión múltiple con dos variables independientes (tamaño de familia y renta) y no utilizar la tercer variable (número de posesión de automóviles) para hacer predicciones.

El análisis de regresión es una técnica de dependencia simple y sencilla que puede proporcionar al analista de datos tanto predicción como explicación. El ejemplo previo ha ilustrado los conceptos y procedimientos básicos del análisis de regresión con el fin de desarrollar un conocimiento de la racional y las cuestiones de este procedimiento en su forma más básica. Los siguientes puntos tratan estas cuestiones con más detalle.

Un proceso de decisión para el análisis de regresión múltiple

El proceso de decisión para el análisis de regresión múltiple comienza con la especificación de los objetivos del análisis de regresión, incluyendo la selección de las variables dependientes e independientes. Entonces se procede a diseñar el análisis de regresión con la consideración de factores tales como el tamaño muestral y la necesidad de transformaciones de variables. Una vez formulado el modelo de regresión, se contrastan en primer lugar los supuestos subyacentes al análisis de regresión para las variables individualmente. Si se cumplen todos los supuestos, entonces se estima el modelo. Una vez que se obtienen los resultados, se lleva a cabo el análisis de diagnóstico para asegurar que el modelo global cumple los supuestos de regresión y que ninguna observación tiene una influencia indebida sobre los resultados. El siguiente paso es la interpretación del valor teórico de la regresión, donde se examina el papel jugado por cada variable independiente en la predicción de la medida dependiente. Finalmente, los resultados se validan para asegurar la generalidad de su población.

PROCESO DE DECISIÓN PARA EL ANÁLISIS DE REGRESIÓN MÚLTIPLE



Objetivos de la regresión múltiple

El análisis de regresión múltiple, una forma de modelo lineal general, es una técnica estadística de análisis multivariante utilizada para examinar las relaciones entre una única variable criterio y un conjunto de variables independientes. El punto de partida necesario del análisis multivariante, como en todas las técnicas multivariantes, es el problema a investigar. La flexibilidad y la capacidad de adaptación de la regresión múltiple permite utilizarlos con casi cualquier relación de dependencia. En la selección de la aplicación adecuada del análisis de regresión, se deben considerar tres asuntos fundamentales:

1. ***La conveniencia del programa de investigación***
2. ***La especificación de una relación estadística***
3. ***La selección de las variables dependientes e independientes***

Problemas de investigación adecuados para la regresión

Las crecientes aplicaciones de la regresión múltiple se agrupan en dos amplias clases de problemas de investigación:

a) Predicción
b) Explicación

Estos problemas de investigación no son mutuamente excluyentes y se puede llevar a cabo una aplicación del análisis de regresión múltiple para cualquiera de los dos tipos de problemas de investigación.

a) Predicción con Regresión Múltiple

Un propósito fundamental de la regresión múltiple es la predicción de una variable criterio con un conjunto de variables independientes. Al hacerlo, la regresión múltiple cumple uno de dos objetivos. El primer objetivo es maximizar la potencia conjunta de predicción de las variables independientes tal y como se representan en el valor teórico. Esta combinación lineal de variables independientes se construye de tal forma que se convierta en un predictor óptimo de la variable criterio.

La regresión múltiple proporciona un medio objetivo de evaluar el poder predictivo de un conjunto de variables independientes. En aplicaciones centradas en este objetivo, el análisis está principalmente orientado a conseguir la máxima predicción. La regresión múltiple proporciona muchas opciones tanto en la forma como en la especificación de las variables independientes que puedan modificar el valor teórico para aumentar su poder predictivo. Muchas veces la predicción se maximiza a expensas de la interpretación. Un ejemplo de una variante de análisis de regresión, el análisis de series temporales, en el cual el único propósito es predecir y la interpretación de los resultados es útil sólo como un medio de incrementar la precisión predictiva. En otras situaciones, la precisión predictiva es crucial para asegurar la validez del conjunto de variables independientes, teniendo en cuenta la ulterior interpretación del valor teórico. Las medidas de precisión predictiva y los tests estadísticos se forman en relación con la significación del poder predictivo que pueda obtenerse. En todos los casos, tanto si la predicción es o no el objetivo principal, el análisis de regresión debe conseguir niveles aceptables de precisión predictiva para justificar su aplicación. Se debe asegurar tener en cuenta tanto la significación práctica como la estadística.

La regresión múltiple puede también conseguir un segundo objetivo de comparación de dos o más conjuntos de variables independientes para averiguar el poder predictivo de cada valor teórico. Ilustrativo de una aproximación de modelización confirmatoria, este uso de la regresión múltiple se centra en la comparación de resultados entre dos o más alternativas o modelos en competencia. El objetivo principal de este tipo de análisis es el poder predictivo relativo entre modelos, aunque en cualquier situación la predicción del modelo elegido debe demostrar tanto significación práctica como estadística.

b) Explicación con Regresión Múltiple

La regresión múltiple proporciona también un medio de evaluar objetivamente el grado y características de la relación entre las variables dependiente y las variables independientes al formar el valor teórico. Las variables independientes, además de su predicción conjunta de la variable dependiente, pueden considerarse también por su contribución individual al valor teórico y a sus predicciones. La interpretación del valor teórico puede tomarse desde alguna de las tres perspectivas:

- ✓ ***la importancia de las variables independientes***
- ✓ ***los tipos de relaciones encontradas***
- ✓ ***las interrelaciones entre las variables independientes.***

La interpretación más directa del valor teórico de la regresión es una determinación de la importancia relativa de cada variable independiente en la predicción de la medida independiente. En todas las aplicaciones, la selección de variables independientes se basaría en sus relaciones teóricas con la variable dependiente. El análisis de regresión proporciona un medio de evaluar objetivamente la magnitud y dirección (positiva o negativa) de cada relación con la variable independiente. El carácter de la regresión múltiple es la evaluación simultánea de relaciones entre cada variable independiente y las medidas de la dependiente. Al realizar esta evaluación simultánea, se determina la importancia relativa de cada predictor.

Además de evaluar la importancia de cada variables, la regresión múltiple permite la evaluación de la naturaleza de las relaciones entre las variables independientes y la variable dependiente. La relación supuesta es una asociación lineal basada en correlaciones entre las variables independientes y la variable

dependiente. Pero también se disponen de transformaciones para evaluar si existen otros tipos de relación, particularmente las relaciones curvilineales.

Finalmente, la regresión múltiple proporciona una idea de las relaciones entre las variables independientes en sus predicciones de la variable dependiente. Estas interpretaciones son importantes por dos razones. En primer lugar, la correlación entre las variables independientes puede hacer que algunas variables sean redundantes en su esfuerzo predictivo. Como tal, no son necesarias para producir una predicción óptima. No se trata de reflejar sus relaciones individuales con la variable criterio sino que indica que en un contexto multivariante, no son necesarias si se emplea otro conjunto de variables independientes para explicar la varianza. Las interrelaciones entre las variables pueden extenderse no sólo a su poder predictivo sino también a las interrelaciones entre sus efectos estimados. Esto se ve mejor cuando el efecto de una variable independiente es contingente con otra variable independiente. La regresión múltiple proporciona un diagnóstico que puede determinar si existen tales efectos basados en razones empíricas o teóricas. Las indicaciones con un alto grado de interrelaciones (multicolinealidad) entre las variables independientes pueden sugerir el uso de las escalas sumadas.

Especificación de la relación estadística
Una regresión múltiple es apropiada cuando el interés está centrado en una relación estadística, no funcional.
Una relación funcional calcula un valor exacto mientras que una relación estadística estima un valor medio.

Selección de variables dependientes e independientes

El "éxito" final de cualquier técnica, incluyendo las regresiones múltiples, comienza con la selección de las variables que se van a utilizar en el análisis. Dado que la regresión múltiple muestra una relación de dependencia, se debe especificar qué variable se usa como criterio y qué variables se usan como predictor. La selección de ambos tipos de variables debería basarse principalmente en fundamentos conceptuales o teóricos.

La selección de una variable criterio está muchas veces dictada por el problema de la investigación. Pero en muchos casos, se debe ser consciente del **error de medida**, especialmente en la variable independiente. El error de medida se refiere al grado en que la variable es una medida precisa y consistente del concepto que está siendo estudiado. Si la variable que se utiliza como dependiente tiene una error sustancial de medida, entonces incluso las mejores variables independientes pueden ser incapaces de conseguir niveles aceptables de precisión predictiva. El error de medida puede venir de diversas fuentes que van desde errores en la entrada de datos a la imprecisión en la medición, pasando por la incapacidad de los encuestados a proporcionar información precisa. El impacto del error de medida es añadir "ruido" a las variables medidas u observadas. El error de medida que es problemático puede ser abordado mediante el uso de las escalas aditivas. Siempre se debe tratar de obtener la mejor medida de las variables dependientes e independientes, basadas ambas en factores empíricos y conceptuales.

El supuesto más problemático en la selección de variables independientes es el **error de especificación**, que hace referencia a la inclusión de variables irrelevantes o a la omisión de variables relevantes del conjunto de variables independientes. Aunque la inclusión de una variable irrelevante no sesgue los resultados de la otras variables independientes, tiene cierto impacto sobre ellos:

- ✓ ***Reduce la parsimonia del modelo, que puede ser crítica en la interpretación de los resultados***
- ✓ ***Las variables adicionales pueden enmascarar o desplazar los efectos de variables más útiles.***
- ✓ ***Las variables adicionales pueden hacer que las contrastaciones de la significación estadística de las variables independientes sean menos precisas y reduzcan la significación estadística y práctica del análisis.***

Por otro lado, la exclusión de variables relevantes puede sesgar seriamente los resultados y afectar negativamente cualquier interpretación de ellos. Cuando existe correlación entre las variables incluidas y las omitidas, los efectos de las variables incluidas pueden verse sesgados en la medida en que están correlacionados con las variables omitidas. Cuanto mayor sea la correlación, mayor será el sesgo. Los efectos estimados para las variables incluidas representan ahora no sólo sus efectos reales sino también los efectos que las variables incluidas comparten con las variables omitidas. Esto puede llevar a serios problemas en la interpretación de los modelos y en la evaluación de la significación estadística y práctica.
Se debe guardar especial cuidado en la selección de las variables para evitar ambos tipos de errores de especificación. Quizá los mayores problemas consistan en la omisión de las variables relevantes, dado

que los efectos de las variables no pueden evaluarse sin su inclusión. Esto intensifica la necesidad de un soporte práctico y teórico de todas las variables incluidas o excluidas en un análisis de regresión múltiple.

Diseño de la investigación en el análisis de la regresión múltiple

En este punto se deben considerar cuestiones como el tamaño muestral, la naturaleza de las variables independientes y la posible creación de nuevas variables para representar las especiales relaciones entre las variables dependientes e independientes. Al hacerlo, debe mantener siempre el criterio de significación práctica y estadística. La capacidad de las regresiones múltiples para realizar muchos tipos de investigaciones se ve enormemente influenciada por los supuestos del diseño que se detallan a continuación

Tamaño Muestral

El tamaño muestral es quizá el elemento aislado más influyente bajo control del analista de datos en el diseño del análisis. Los efectos del tamaño muestral se ven más directamente en la potencia estadística del test de significación y la generalización del resultado.

Potencia estadística y tamaño muestral

El tamaño muestral tiene un impacto directo en la conveniencia y la potencia estadística de la regresión múltiple. Muestras pequeñas, habitualmente caracterizadas por tener menos de 20 observaciones, son apropiadas solo para el análisis de regresión simple con una única variable independiente. Incluso en estas situaciones, sólo se pueden detectar relaciones muy fuertes con cierto grado de certidumbre. De la misma forma, las muestras muy grandes, de 1.000 observaciones o más, hacen los tests de significación estadística demasiado sensibles, indicando que casi cualquier relación es estadísticamente significativa.
Con muestras muy grandes se debe asegurar que el criterio de significación práctica se cumpla a la vez que la significación estadística.

La potencia de la regresión múltiple se refiere a la probabilidad de detectar como estadísticamente significativo un coeficiente de regresión para un nivel de significación especificado y un tamaño de muestra específico.
El tamaño muestral tiene un impacto directo y cuantificable sobre la potencia.

La siguiente Tabla ilustra la interacción entre el tamaño muestral, el nivel de significación (alfa) elegido y el número de variables independientes para detectar un R^2 significativo.

Tamaño muestral	Nivel de significación (alfa) = 0.01 Número de variables independientes				Nivel de significación (alfa) = 0.05 Número de variables independientes			
	2	**5**	**10**	**20**	**2**	**5**	**10**	**20**
20	45	56	71	No aplicable	39	48	64	No aplicable
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1000	1	2	2	3	1	1	2	2

Tabla: Mínimo R^2 que se puede encontrar estadísticamente significativo con una potencia de 0.8 para diferentes variables independientes y tamaños muestrales

Los valores de la tabla son el mínimo R^2 que el tamaño muestral especificado detectará como estadísticamente significativo y el nivel alfa especificado con un probabilidad (potencia) de 0.80. Por ejemplo, si se emplean 5 variables independientes, se especifica el nivel de significación de 0.05 y se está satisfecho al detectar un R^2 del 80% de las veces que ocurre (correspondiente a una potencia = 0.8), una muestra de 50 encuestados detectará valores de R^2 del 23% y superior. Si la muestra aumenta a 100 encuestados, entonces se detectarán valores del R^2 del 12% o superiores. Pero si los 50 encuestados es todo lo que se tiene y se quiere un nivel de significación del 0.01, se detectarán valores de R^2 sólo por encima del 29%. En estos casos se debería considerar el papel del tamaño muestral en la contrastación de la significación antes de la recogida de datos. Si se esperan relaciones débiles, pueden hacerse juicios contrastados en la medida en que el tamaño de muestra necesario detecte razonablemente las relaciones, si existen. Por ejemplo, en la tabla se demuestra que los tamaños de muestra de 100 detectarán claramente valores R^2 bajos (del 10% al 15%) por encima de 10 variables independientes y un nivel de significación de 0.05. Si embargo, si el tamaño muestral es menor de 50 observaciones en estas situaciones, puede detectarse el doble de R^2.

Generalización y tamaño muestral

El tamaño muestral también afecta a la generalización de los resultados, en función del ratio de observaciones sobre las variables independientes. Una norma general es que ese ratio nunca debería

caer por debajo de 5, lo que significa que existirán 5 observaciones para cada variable independiente presente en el valor teórico. Cuando este nivel se alcance los resultados deberían ser generalizables si el tamaño muestral es representativo.

Creación de Variables adicionales

La relación básica representada en la regresión múltiple es la asociación *lineal* entre variables dependientes e independientes *métricas* basada en la correlación momento-producto. Muchas veces se enfrenta el problema de incorporar datos no métricos, tales como género u ocupación, en una ecuación de regresión. Para esos casos deben crearse nuevas variables mediante transformaciones de los datos. La transformación de los datos ofrece un medio de modificar tanto las variables dependientes como las independientes por una de estas dos razones:

1. *Mejorar o modificar la relación entre las variables dependientes ó independientes*
2. *Permitir el uso de variables no métricas en el valor teórico de la regresión.*

Supuestos en el análisis de regresión múltiple

Valoración de las variables individuales frente al valor teórico

Los supuestos subyacentes a la regresión múltiple se aplican tanto a las variables individuales (dependientes e independientes) como a la relación global. Se deben evaluar los supuestos no sólo de las variables individuales sino del valor teórico en sí mismo. Este punto se centrará en el examen del valor teórico y su relación con la variable dependiente en el cumplimiento de los supuestos de la regresión múltiple.

Los supuestos a examinar son los siguientes:

a) Linealidad del fenómeno

b) Varianza constante del término de error

c) Independencia de los términos de error

d) Normalidad de la distribución del término de error

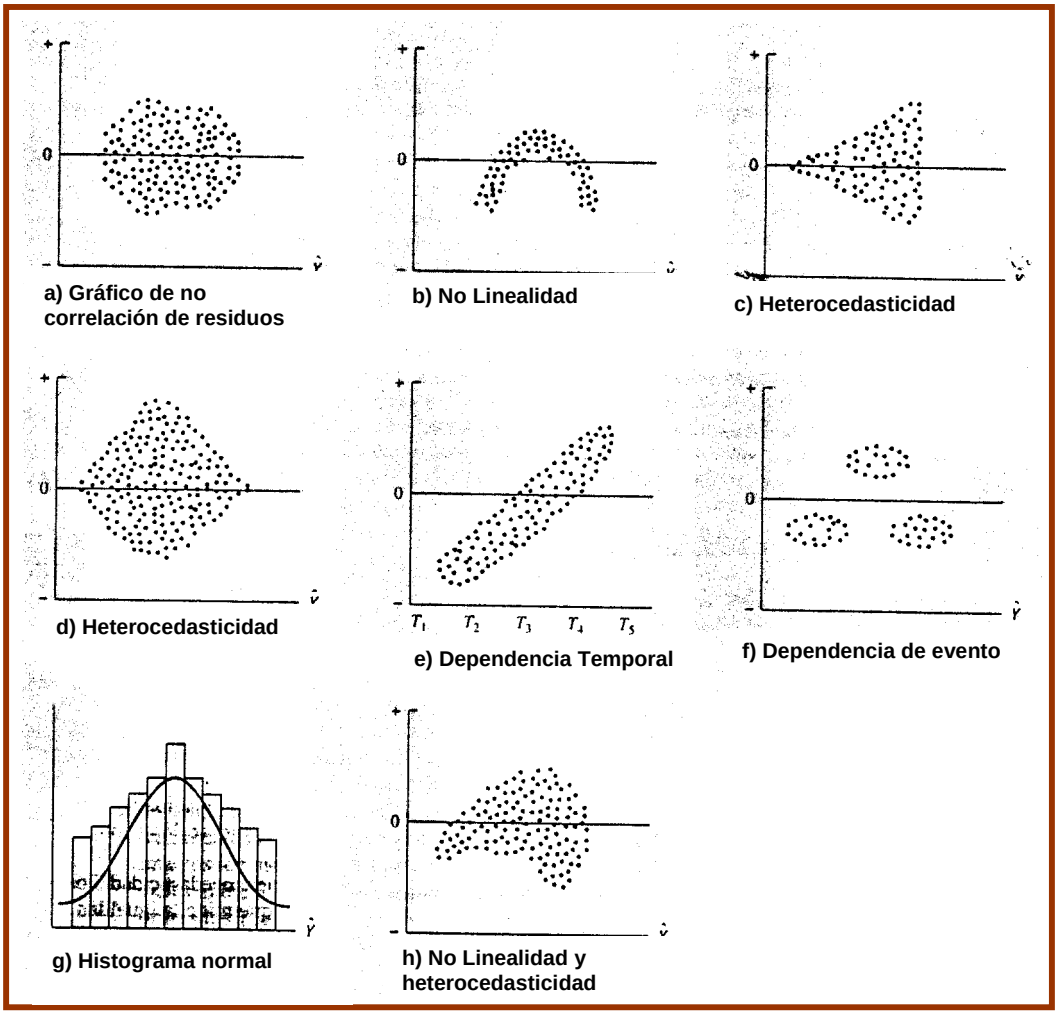
La medida principal del error de predicción del valor teórico es el **residuo** –la diferencia entre los valores observados y las predicciones de la variable criterio. Los gráficos de residuos y de las variables independientes o de las predicciones constituyen el método básico de identificación de los incumplimientos de los supuestos para el conjunto de la relación. Cuando se examinan los residuos, se recomienda cierta forma de estandarización, con el fin de hacer los residuos directamente comparables. El método más ampliamente utilizado es **el residuo basado en la *t* de Student**.

Su valor corresponde a valores t. Esta correspondencia facilita la evaluación de la significación estadística de residuos particularmente elevados.

El gráfico de residuos más habitual se forma con los residuos (r_1) frente a los valores de la predicción de la variable dependiente (Y_i). Para un modelo de regresión simple, se pueden trazar los residuos respecto de las variables dependientes o independientes, dado que están directamente relacionados. En la regresión múltiple, sin embargo, sólo los valores dependientes pronosticados representan el efecto total del valor teórico de la regresión. Por tanto, a no ser que el análisis residual pretenda concentrarse en una sola variable, se usan las variables dependientes pronosticadas.

La **Figura-Análisis gráfico de los residuos-** contiene unos cuantos gráficos de residuos que muestran los supuestos básicos discutidos a continuación.

Figura – Análisis Gráficos de los Residuos



Un gráfico de especial interés es el caso (a)-Gráfico de no-correlación de residuos. Los gráficos de no-correlación de los residuos se distribuyen aleatoriamente, con una dispersión relativamente igual a cero y una tendencia no muy fuerte a que sea mayor o reducidos de la variable independiente.

a) Linealidad del fenómeno

La linealidad de la relación entre variables dependientes o independientes representa el grado de cambio en la variable dependiente asociado con la variable independiente. El coeficiente de regresión es constante a lo largo del rango de valores de la variable independiente. El concepto de correlación está basado en la relación lineal, siendo por tanto un supuesto crítico del análisis de regresión. La figura (b) muestra una forma típica de residuos que indican la existencia de una relación no lineal no representada en el modelo habitual. Cualquier modelo curvilineal de los residuos indica que la acción correctiva aumentará tanto la precisión predictiva del modelo como la validez de los coeficientes estimados.

En la regresión múltiple con más de una variable independiente, el examen de residuos mostraría los efectos combinados de todas las variables independientes, pero no se puede examinar el efecto de cualquier variable independiente separadamente en un gráfico de residuos. Para hacerlo se utiliza lo que se denomina gráfico de regresión parcial que muestra la relación de una única variable independiente con relación a otra variable dependiente. Difiere del gráfico de residuos que se acaba de mencionar en que la línea que atraviesa la nube de puntos, que era horizontal, tendrá ahora una pendiente positiva o negativa dependiendo si el coeficiente de regresión para esa variable independiente es positivo o negativo.

b) Varianza constante del término de error

La presencia de varianzas desiguales (heterocedasticidad) es uno de los supuesto que se incumple más habitualmente. El diagnóstico se realiza mediante gráficos de residuos o test estadísticos simples. El gráfico de los residuos (basado en la t de Student) frente a los valores de la variable dependiente se compone con el gráfico de no-correlación de residuos –Figura (a)- y muestra una forma consistente si la varianza no es constante. Quizá la forma más común es la triangular, en cualquier dirección –Figura (c)-. Puede esperarse una forma de diamante –Figura (d)- en el caso donde se espera más variación en los valores intermedios que en los extremos. Muchas veces, muchos incumplimientos ocurren simultáneamente, tal como la no-linealidad y la heterocedasticidad mostrada en la –Figura (h)-.
Cada programa estadístico de computador tiene test de heterocedasticidad. Por ejemplo, SPSS ofrece el test de Levane de homogeneidad de varianza, que mide la igualdad de varianzas para un único par de varianzas. Su uso es particularmente recomendable porque es el que menos queda afectado por desviaciones de la normalidad, otro de los problemas que ocurren con frecuencia.

c) Independencia de los términos de error

En la regresión se supone que cada variable predictor es independiente. Con ello se quiere decir que el valor de la predicción no está relacionado con cualquier otra predicción; esto es, no están ordenados por otra variable. Para identificar este hecho, se utiliza el gráfico de residuos respecto a cualquier posible varianza secuencial. Si los residuos son independientes, la forma puede parecer aleatoria y similar al gráfico de no-correlación. Los incumplimientos quedarán identificados por una forma consistente de los residuos. La –Figura (e)- representa un gráfico de residuos que muestra una asociación entre los residuos y tiempo, una variable de secuencia habitual. Otra forma habitual se muestra en la –Figura (f)-. Esta forma ocurre cuando las condiciones básicas del modelo cambian pero no se incluyen en el modelo.

d) Normalidad de la distribución del término de error

Quizá el incumplimiento de supuestos más frecuentes la no-normalidad de las variables independientes y dependientes, o ambos. El diagnóstico más simple para el conjunto de variables predictor en la ecuación es un histograma de residuos, donde se puede comprobar visualmente si la distribución se aproxima a la normal.

El análisis de residuos, bien con los gráficos de residuos bien con test estadístico, proporciona un conjunto simple pero potente de instrumentos analíticos para examinar la conveniencia del modelo de regresión.

Estimación del modelo de regresión y valoración

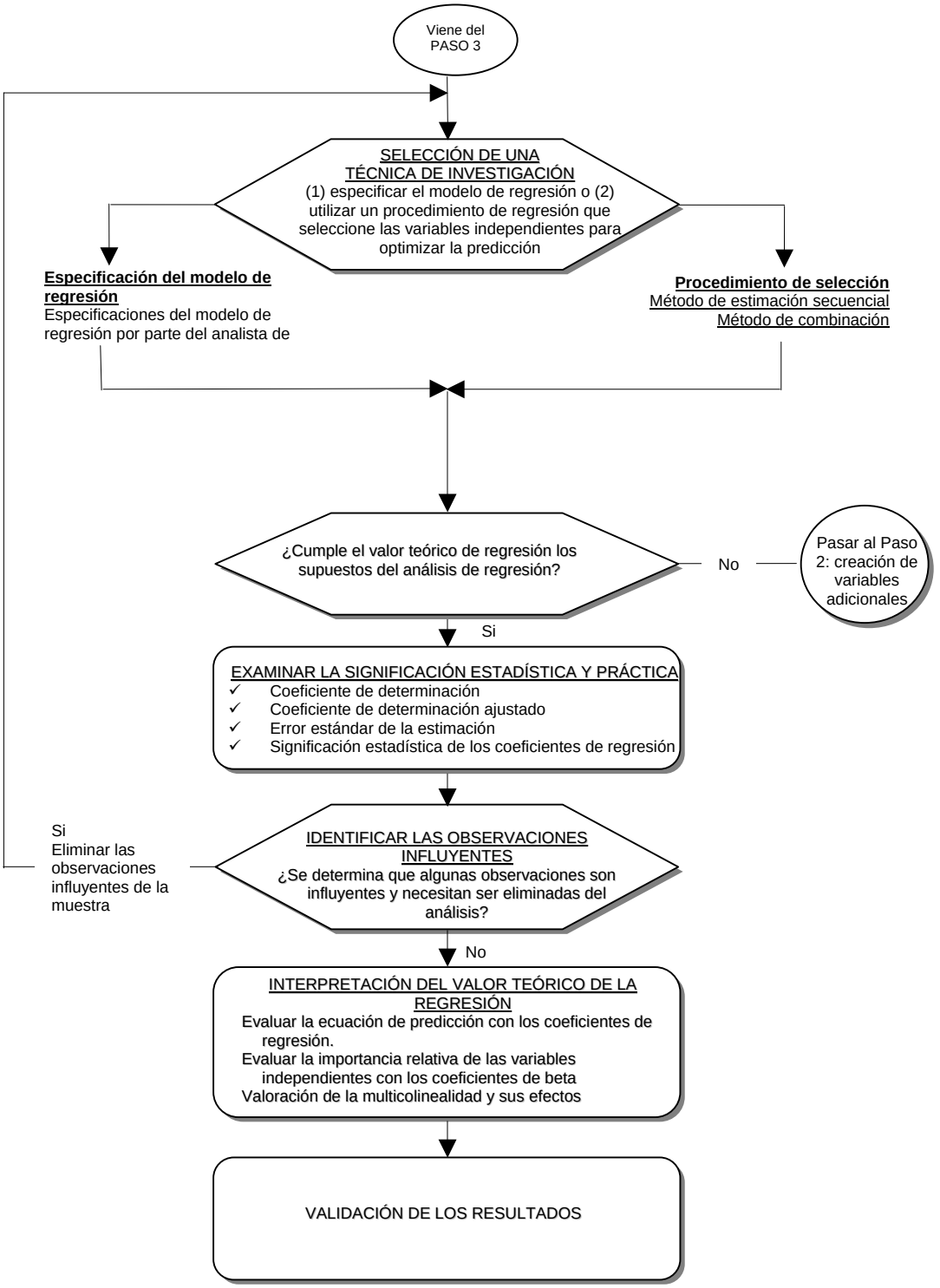
Habiendo especificado los objetivos del análisis de regresión, seleccionado las variables dependientes e independientes, enfrentados los resultados del diseño de la investigación y evaluadas las variables a la hora de cumplir los supuestos de la regresión, se está preparado para estimar el modelo de regresión y evaluar la precisión predictiva conjunta de las variables independientes. A este nivel se deben lograr tres tareas básicas:

1. **Seleccionar un método para especificar el modelo de regresión a estimar.**

2. **Evaluar la significación estadística del modelo conjunto en la predicción de la variable criterio.**

3. **Determinar si cualquiera de las observaciones ejerce una indebida influencia sobre los resultados.**

PROCESO DE DECISIÓN PARA EL ANÁLISIS DE REGRESIÓN MÚLTIPLE
PARTE II



Aproximaciones generales a la selección de variables

En la mayoría de los casos de regresión múltiple, se tiene un número posible de variables independientes entre las cuales elegir para incluirlas en la ecuación de regresión. A veces el conjunto de variables independientes puede estar muy definido y el modelo de regresión se usa esencialmente en una **aproximación confirmatoria**. En otros caso, se puede elegir entre el conjunto de variables independientes. Existen varias aproximaciones para la búsqueda del mejor modelo de regresión: **métodos de búsqueda secuencial** y **procesos combinatorios**.

- **Especificación confirmatoria**

  La más simple aunque quizá más exigente aproximación de especificación del modelo de regresión es emplear una perspectiva confirmatoria en la cual se especifica por completo el conjunto de variables independientes a incluir. En este caso se tiene control total sobre la selección de la variable, por tal motivo se requiere la absoluta seguridad de que el conjunto de variables consigue la máxima predicción mientras mantiene un modelo de parsimonia (grado en que modelo logra la calidad del ajuste para cada coeficiente estimado, con el objetivo de maximizar la cantidad de ajuste por coeficiente estimado y evitar "sobre-ajustar" el modelo con coeficientes adicionales que sólo consiguen pequeñas ganancias en el ajuste del modelo).

- **Métodos de búsqueda secuencial**

  Los métodos de búsqueda secuencial tienen en común la aproximación general de estimación de las ecuaciones de regresión con un conjunto de variables y a continuación añadir o eliminar selectivamente variables hasta que se consiga alguna medida criterio conjunta. Esta aproximación proporciona un método objetivo de selección de variables que maximizan la predicción con el número más pequeño de variables empleadas. En cada aproximación, se valoran las variables individualmente en función de su contribución a la predicción de la variable dependiente y se añaden o eliminan según su contribución relativa.

- **Métodos combinatorios**

  Los métodos combinatorios son fundamentalmente un proceso de búsqueda generalizada a lo largo de todas las combinaciones posibles de variables independientes. El procedimiento más conocido en la **regresión parcial combinando variables**, que es exactamente lo que su nombre indica. Se examinan todas las combinaciones posibles de las variables independientes para identificar el conjunto de variables que mejor se ajusta. Por ejemplo, en un modelo con diez variables independientes, existen 1.201 regresiones posibles (una ecuación con una única constante, 10 ecuaciones con una única variable independiente, 45 ecuaciones con todas las combinaciones posibles de dos variables, etc.). Con procedimientos de estimación informáticos, este proceso se puede gestionar incluso para problemas muy grandes, identificando la mejor ecuación de regresión conjunta para cualquier número de medidas de ajuste predictivo.

  Perspectiva de las aproximaciones de la selección de modelos
  Independientemente del modelo seleccionado, el criterio más importante es el conocimiento sustantivo del analista de datos de la situación, que es lo que determina las variables que se van a incluir así como los signos esperados y la magnitud de sus coeficientes. Sin este conocimiento, la regresión resultante puede tener una elevada precisión predictiva sin relevancia teórica o gerencial.

Contrastación del cumplimiento de los supuestos de regresión

Con las variables independientes seleccionadas y los coeficientes de regresión estimados, se debe evaluar el modelo estimado a la hora de cumplir los supuestos subyacentes de la regresión múltiple. Las variables individuales deben cumplir los supuestos de linealidad, varianza constante, independencia y normalidad. Además de las variables individuales, el valor teórico de la regresión debe también cumplir estos supuestos.

Identificación de observaciones influyentes

Hasta ahora e desarrollo se ha centrado en la identificación de pautas generales en el conjunto de observaciones. Aquí se desviará la atención a las observaciones individuales, con e objetivo de encontrar aquellas que caen fuera de las pautas generales del conjunto de datos o que ejercen una fuerte influencia en los resultados de la regresión. Estas observaciones no son necesariamente malas en el sentido que

deban ser omitidas. En muchos casos representan los elementos diferenciadores del conjunto de datos. Sin embargo, deben ser previamente identificadas y evaluado su impacto.

Las observaciones influyentes se clasifican en tres casos:

1. Atípicos
2. Puntos de apalancamiento
3. Influyentes

1. Los "***atípicos***" son observaciones que tienen grandes valores residuales y pueden identificarse sólo con respecto a un modelo de regresión específico.

2. Los "***puntos de apalancamiento***" son observaciones diferentes del resto de las observaciones de los valores de las variables independientes. Su impacto es particularmente destacado en los coeficientes estimados de una o más variable predictor.

3. Las "***observaciones influyentes***" en sentido amplio, incluyen todas las observaciones que tienen un efecto desproporcionado sobre los resultados de la regresión. Las observaciones influyentes no sólo incluyen potencialmente atípicos y puntos de apalancamiento, sino que pueden incluir otras observaciones. Por otro lado, no todos los atípicos y puntos de apalancamiento son necesariamente observaciones influyentes.

   La necesidad de un estudio adicional de los puntos de apalancamiento y los influyentes se pone de manifiesto cuando se ve la sustancial medida en la que la generalización de los resultados y las conclusiones sustantivas (la importancia de las variables, nivel de ajuste, etc.) pueden modificarse por un número relativamente pequeño de observaciones. Sean "buenas" (acentuando los resultados) o "malas" (cambiando sustancialmente los resultados), estas observaciones deben identificarse para evaluar su impacto. Influyentes, atípicos y puntos de apalancamiento se basan en alguna de las cuatro condiciones siguientes:

   a) Un error en la entrada de observaciones o datos.
   b) Una observación válida aunque excepcional que es explicable por una situación extraordinaria
   c) Una observación excepcional sin una explicación plausible.
   d) Una observación ordinaria en sus características individuales pero excepcional en su combinación de características.

Interpretación del valor teórico de la regresión

La interpretación del valor teórico de la regresión implica evaluar no sólo el modelo de regresión que se estimó sino también el potencial de variables independientes que se omitieron si se empleó una aproximación combinatorial o de búsqueda secuencial. En estas aproximaciones, la multicolinealidad puede afectar sustancialmente a las variables incluidas en última instancia en el valor teórico de la regresión. Además de evaluar los coeficientes estimados, se debe evaluar también el impacto potencial de las variables omitidas para asegurar que la significación práctica se evalúa a la vez que la significación estadística.

Utilización de los coeficientes de regresión

Los coeficientes de regresión estimados se usan para calcular los valores de la predicción para cada observación y para expresar el cambio esperado de la variable dependiente para cada unidad de cambio en las variables independientes. Además de hacer la predicción, es deseable saber qué variable independiente es la más útil en la predicción de la variable dependiente.

Estandarización de los coeficientes de regresión: los coeficientes beta

Si cada una de las variables predictor ha sido **estandarizada** antes de estimar la ecuación de regresión, se presentarían diferentes coeficientes de regresión. Los coeficientes resultantes de los datos estandarizados se denominan **coeficientes beta**. Su valor reside en que eliminan el problema de tratar con diferentes unidades de medida y reflejan el impacto relativo sobre la variable criterio de un cambio en una desviación estándar de cada variable. Cuando se tiene una unidad común de medida, se puede determinar qué variable es la más influyente.

Se deben tener en cuenta tres precauciones cuando se utilizan los coeficientes beta:

a) Deben utilizarse como guía de la importancia relativa de las variables individuales dependientes únicamente cuando la colinealidad es mínima.
b) Los valores beta pueden interpretarse sólo en el contexto de las otras variables de la ecuación.
c) Los niveles de las variables afectan al valor beta.

Los coeficientes beta se deben utilizar sólo como guía de la importancia relativa de las variables predictor incluidas en la ecuación y sólo en el rango de valores para el que realmente existe una muestra de datos.

Evaluación de la multicolinealidad

Un supuesto clave en la interpretación del valor teórico de la regresión es la correlación entre las variables predictor. Se trata de un problema de datos, no de un problema de especificación del modelo. La situación ideal sería tener una cantidad de variables independientes que estuvieran altamente correlacionadas con la variable dependiente, pero con poca correlación entre sí. Sin embargo, en la mayoría de las situaciones, especialmente las situaciones que incluye datos de respuesta de consumidores, habrá algo de multicolinealidad. En otras ocasiones, como las de la utilización de variables ficticias para representar variables no métricas o términos polinomiales para efectos no lineales, se crean situaciones con alta multicolinealidad. La tarea es: (1) valorar el grado de multicolinealidad y (2) determinar su impacto en los resultados y las soluciones pertinentes para cada necesidad.

Los efectos de la multicolinealidad

Se puede clasificar los efectos de la multicolinealidad en términos de explicación y estimación. Los efectos sobre la explicación conciernen principalmente a la capacidad del procedimiento de regresión y a la capacidad del analista de datos para representar y comprender los efectos de cada variable independiente en el valor teórico de regresión. Conforme ocurre la multicolinealidad, el proceso para la separación de los efectos de los individuos es cada vez más difícil. En primer lugar, limita el tamaño del coeficiente de determinación y hace más difícil añadir una predicción explicatoria extra con variables adicionales. En segundo lugar, hace difícil determinar la contribución de cada variable debido a que los efectos de las variables independientes son mixtos y se confunden. La multicolinealidad tiene como resultado porciones más grandes de la varianza compartida y niveles más bajos de varianza única a partir de los cuales se pueden determinar los efectos de las variables independientes individuales.

Además de los efectos sobre la explicación, la multicolinealidad puede tener efectos sustantivos sobre la estimación de los coeficientes de regresión y sus pruebas de significación estadística. En primer lugar, el caso extremo de la multicolinealidad en la cual dos o más variables están perfectamente correlacionadas, que se denomina **singularidad**, impide la estimación de cualquier coeficiente. En este caso, la singularidad tiene que ser eliminada para que la estimación de los coeficientes pueda proceder. Incluso si la multicolinealidad no es perfecta, altos grados de multicolinealidad pueden tener como resultado la incorrecta estimación de los coeficientes de regresión.

La identificación de la multicolinealidad

En múltiples análisis de regresión, la evaluación de la multicolinealidad se debe realizar en dos pasos: (1) identificación de la magnitud de la colinealidad y (2) la evaluación del grado en que los coeficientes estimados se ven afectados.

El medio más simple de identificar la colinealidad es un examen de la matriz de correlación de las variables independientes. La presencia de una elevada correlación (generalmente de 0.9 en adelante) es la primera indicación de una elevada colinealidad. La ausencia de elevados valores de correlación no asegura la falta de colinealidad. La colinealidad puede deberse a los efectos combinados de dos o más variables independientes.

Dos de las medidas más comunes para evaluar la colinealidad de parejas o de múltiples variables son:
1. El valor de tolerancia
2. Su inverso, el factor de inflación de la varianza (VIF)

Estas medidas indican el grado en el que cada variable independiente se explica por otras variables independientes. En términos simples, cada variable independiente se convierte en una variable criterio y se realiza la regresión con el resto de las variables independientes. La tolerancia es la cantidad de variabilidad de las variables independientes seleccionadas no explicadas por el resto de las variables independientes. Por lo tanto, un valor de tolerancia reducido (o elevados valores VIF) denotan una elevada colinealidad. Cada analista de datos debe determinar el grado de colinealidad que aceptará, en la medida en que los límites por defecto o los recomendados pueden aceptar todavía una colinealidad sustancial.

Remedios para la multicolinealidad

Los remedios para la multicolinealidad se extienden desde una modificación del valor teórico de regresión hasta el uso de los procedimientos de estimación especializada Una vez determinado el grado de colinealidad, se tienen varias opciones:

- Omitir una o varias variables independientes correlacionadas e identificar otras variables independientes para ayudar con la predicción (evitando la creación de un error de especificación cuando se eliminan una o más variables independientes)
- Utilizar el modelo con las variables correlacionadas solo para predecir (no intentando interpretar los coeficientes de regresión)
- Utilizar las correlaciones simples entre cada variable independiente y cada variable dependiente para entender la relación entre la variable independiente –dependiente
- Utilizar un método más sofisticado de análisis, como una regresión bayesiana o una regresión de componentes principales para obtener un modelo que refleje más claramente los efectos simples de las variables independientes.

Validación de los resultados

El paso final consiste en asegurarse que el modelo represente a la población general (generalización) y que sea apropiada para situaciones en las cuales será utilizada (transferibilidad). La mejor guía es ver en qué medida se ajusta a un modelo teórico o a un conjunto de resultados validados previamente sobre el mismo asunto. En muchos casos, sin embargo, anteriores resultados o la teoría no están disponibles.

Existen aproximaciones empíricas a la validación del modelo.

- **Muestras adicionales o muestras divididas**

  La aproximación más apropiada para la validación empírica es contrastar el modelo de regresión mediante la extracción de una nueva muestra de la población general. Una nueva muestra asegura la representatividad y puede utilizarse en varias formas. En primer lugar, el modelo original puede predecir valores con la nueva muestra, además del ajuste predictivo. En segundo lugar, se puede estimar un modelo separado con la nueva muestra para compararla a continuación con la ecuación original sobre las características de las variables incluidas. Signo, tamaño e importancia relativa de las variables y precisión predictiva. En ambos casos se determina la validez del modelo original comparándolo con los modelos de regresión estimados con la nueva muestra.

  Muchas veces la capacidad de recoger nuevos datos está limitada o impedida por factores como los costos, la falta de tiempo o la disponibilidad de los encuestados. En este caso se puede entonces dividir la muestra en dos partes: una sub-muestra de estimación para crear el modelo de regresión y una sub-muestra de validación / duración utilizada para “contrastar” la ecuación.

  Tanto si se extrae una nueva muestra como si no, es probable que existan diferencias entre el modelo original y otros esfuerzos de validación ya que ningún modelo de regresión, a menos que se estime del conjunto de la población, es el modelo final y absoluto.

- **Cálculo del estadístico PRESS**

  Una aproximación alternativa a la obtención de muestras adicionales con el fin de validar el modelo es emplear la muestra original de forma especializada mediante el **cálculo del estadístico PRESS,** una medida similar a R^2 utilizada para evaluar la precisión predictiva del modelo de regresión estimado. Omite una observación en la estimación del modelo de regresión y predice a continuación las observaciones omitidas con el modelo estimado. De esta forma, la observación no puede afectar a los coeficientes del modelo utilizado para calcular su valor predictivo. El procedimiento se aplica otra vez, omitiendo otra observación, estimando un nuevo modelo y realizando la predicción. Pueden sumarse los residuos de la observaciones para obtener una medida conjunta del ajuste predictivo.

- **Predicción con el modelo**

  Las predicciones del modelo siempre pueden realizarse aplicando el modelo estimado para un nuevo conjunto de valores de las variables independientes y calculando los valores de las variables criterio.

Al hacerlo se deben considerar los siguientes factores:

1. *Cuando se aplica el modelo a una nueva muestra, se debe recordar que las predicciones contienen ahora no sólo las variaciones respecto de la muestra original sino también la muestra nuevamente extraída.*

2. *Asegurarse que las condiciones y relaciones medidas en el momento en que la muestra original fue tomada no han cambiado.*

3. *Utilizar el modelo para estimar dentro del rango de las variables independientes que se encuentran en la muestra y no más allá de dicho rango.*

REGRESIÓN LOGÍSTICA (ANÁLISIS LOGIT)

Sin duda la regresión múltiple es la técnica de dependencia multivariante más utilizada por su capacidad para predecir y explicar las variables métricas. Pero ¿qué decir sobre las variables no métricas?. En este punto se presentará una técnica que se aplica cuando la variable tiene estas características.

La regresión logística cuenta con amplias aplicaciones en situaciones donde el primer objetivo es identificar el grupo al cual un objeto (una persona, una empresa o un producto) pertenece.

Las aplicaciones potenciales incluyen:

- **la predicción de éxitos o fracasos de un nuevo producto**
- **determinar en qué categoría de riego de crédito se encuentra una persona**
- **decidir si un estudiante debe ser admitido en una universidad**
- **determinar el riesgo de que un cliente canjee una recompensa (aplicable a un programa de fidelidad)**

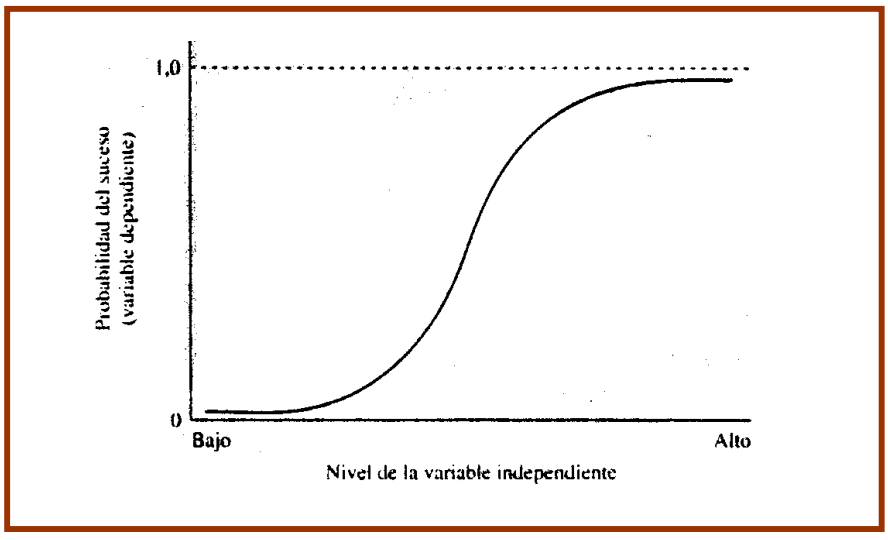
Al intentar elegir una técnica analítica apropiada, algunas veces se encuentra el problema que incluye una variable dependiente categórica y varias variables independientes métricas.

La regresión logística (al igual que el análisis discriminante) son las técnicas apropiadas cuando la variable dependiente es categórica (nominal o no métrica) y las variables independientes son métricas. En muchos casos, la variable dependiente consta de dos grupos o clasificaciones (por ejemplo masculino frente a femenino; ó alto frente a bajo). En otras situaciones, se incluyen más de dos casos, como en una clasificación de tres grupos que comprenda clasificaciones bajas, medias y altas.

La **regresión logística** está restringida en su forma básica a dos grupos. El análisis discriminante tiene la capacidad de tratar tanto dos grupos como grupos múltiples (tres o más).[2]

La regresión logística se diferencia de la regresión múltiple en que predice directamente la probabilidad de ocurrencia de un suceso. Aunque el valor de la probabilidad sea una medida métrica, existen diferencias fundamentales entre la regresión múltiple y la logística. Los valores de la probabilidad pueden ser cualesquiera entre cero y uno, pero el valor predicho debe estar acotado para que caiga en el rango de cero y uno. Para definir una relación acotada entre cero y uno, la regresión logística utiliza una relación supuesta entre las variables dependientes e independientes que recuerda a una curva en forma de S - **Figura RL1-Forma de la relación logística entre las variables dependientes e independientes**.

Figura RL1-Forma de la relación logística entre las variables dependientes e independientes



[2] En este punto se profundizará el desarrollo de regresión logística, ya que es una técnica aplicada con un objetivo de negocio en el caso práctico. En la bibliografía referencial indicada, el lector tendrá la posibilidad de desarrollar el modelo de análisis discriminante.

Para niveles muy bajos de al variable independiente, la probabilidad se aproxima a cero. Según crece la variable independiente, la probabilidad crece a lo largo de la curva, pero como la pendiente empieza a decrecer para cierto nivel de la variable independiente, la probabilidad se acercará a uno sin llegar a excederlo. Como se ha visto en el desarrollo de Regresión Múltiple, los modelos de regresión lineal no permitían captar tal relación, al ser inherentemente no lineal. Además tales situaciones no pueden estudiarse mediante la regresión ordinaria, porque al hacerlo se incumplen varios supuestos, a saber:

1. ***El término de error de una variable discreta sigue una distribución binomial en lugar de la distribución normal, invalidando todos los contrastes estadísticos basados en el supuesto de normalidad.***

2. ***La varianza de una variable dicotómica no es constante.***

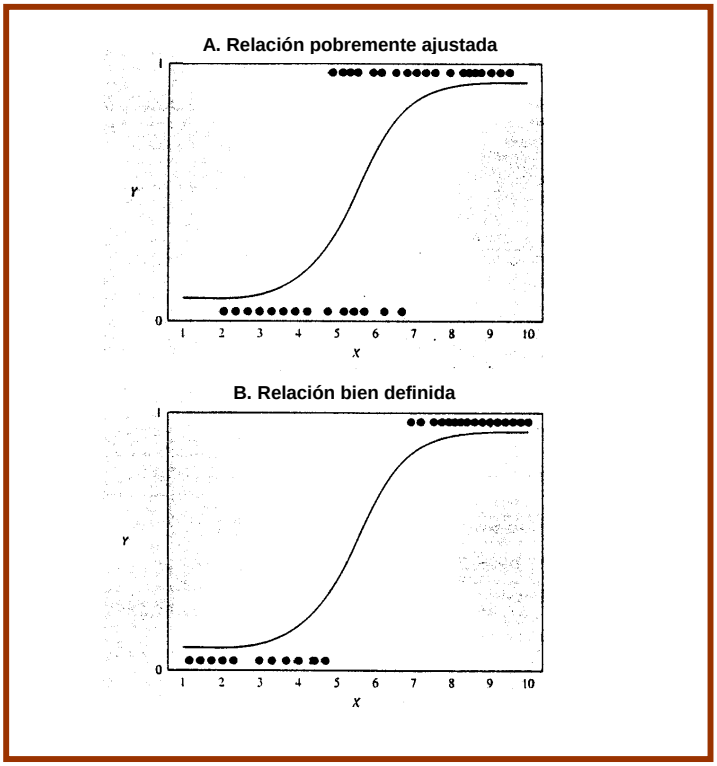
La regresión logística se desarrolló para tratar precisamente estas situaciones. La única relación entre variables dependientes e independientes precisa de una aproximación algo diferente a la estimación, la evaluación de la bondad del ajuste y la interpretación de los coeficientes.

Estimación del modelo de regresión logística

La naturaleza no lineal de la transformación logística requiere que otro procedimiento, el de máxima verosimilitud, se utilice de forma iterativa para encontrar la estimación “más probable” de los coeficientes. Para ello, se usa el valor de la verosimilitud en lugar de la suma de los cuadrados al calcular la medida de ajuste global del modelo.

El modelo logístico tiene la forma concreta de una curva logística. Para estimar el modelo de regresión logística, se ajusta esta curva a los datos reales. La **Figura RL2-Ejemplos de ajuste de la curva logística a datos muestrales** representa dos ejemplos hipotéticos de ajuste de una relación logística a datos muestrales. Los datos reales, que un suceso tenga o no lugar (1 o 0), se representan como observaciones en lo alto o en lo bajo del gráfico. Estos son los sucesos que ocurren para cada valor de la variable independiente (el eje de las X). En la parte A, la curva logística no puede ajustar los datos bien porque hay varios valores de la variable independiente que cuentan tanto con sucesos como no sucesos (esto es, un importante solapamiento de las distribuciones). Sin embargo, en la parte B, existe una relación mucho más definida y la curva logística se ajusta a los datos bastante bien. Este sencillo ejemplo, similar a una nube de puntos entre las variables dependiente e independiente de la regresión con una línea que representa el “mejor ajuste” de la correlación, puede extenderse para incluir múltiples variables independientes como en la regresión.

Figura RL2-Ejemplos de ajuste de la curva logística a datos muestrales



Interpretación de los Coeficientes

Una de las ventajas de la regresión logística es que sólo se necesita saber si un suceso ocurrió (comprar o no, riesgo de crédito o no, riesgo de canje o no) para entonces utilizar un valor dicotómico como la variable dependiente. A partir de este valor dicotómico, el procedimiento predice su estimación de la probabilidad de que el suceso tenga o no lugar. Si la predicción de la probabilidad es mayor que 0.5, entonces la predicción es sí, y no en otro caso. La regresión logística deriva su nombre de la transformación logística utilizada con la variable dependiente. Cuando se utiliza esta transformación, sin embargo, la regresión logística y sus coeficientes tienen un sentido diferente del que se encuentra en la regresión con una variable dependiente métrica.

El procedimiento que calcula el **coeficiente logístico** compara la probabilidad de la ocurrencia de un suceso con la probabilidad de que no ocurra. Este **odds ratio** puede expresarse como:

$$\frac{\text{Prob}_{(\text{evento})}}{\text{Prob}_{(\text{no evento})}} = e^{B_0 + B_1 X_1 + \ldots\ldots B_n X_n}$$

Los coeficientes estimados ($^{B_0 + B_1 X_1 + \ldots\ldots B_n X_n}$) son en realidad medidas de los cambios en el ratio de probabilidades, denominado **odds ratio.** Más aún, están expresados en algoritmos, por lo que necesitaríamos retransformarlos (tomando los valores del anti-logaritmo) de tal forma que se evalúe más fácilmente su efecto sobre la probabilidad. Los programas informáticos lo hacen automáticamente calculando tanto el coeficiente real como el transformado. Utilizar este procedimiento no cambia en modo alguno la forma de interpretar el signo del coeficiente. Un coeficiente positivo aumenta la probabilidad y de lo contrario disminuye.

Si B_i es positivo, su transformación será mayor a 1 y el odds ratio aumentará. Este aumento se produce cuando la probabilidad prevista de ocurrencia de un suceso aumenta y la probabilidad prevista de su no ocurrencia disminuye. Por tanto, el modelo tiene una elevada probabilidad de ocurrencia. De la misma forma si B_i es negativo, su transformación será menor a 1 y el odds ratio disminuirá. Un coeficiente cero indica que no se producen cambios en el odds ratio.

Para representar la relación de una curva en forma de S o logística, los coeficientes deben representar efectivamente relaciones no lineales entre las variables dependientes e independientes. Aunque el proceso de transformación de tomar logaritmos proporciona una linealización de la relación, se debe recordar que los coeficientes representan en realidad diferentes pendientes en la relación entre los valores de la variable independiente. De esta forma puede estimarse la relación en forma de S.

Valoración de la bondad del ajuste del modelo estimado

La regresión logística maximiza la "verosimilitud" de que un suceso tenga lugar. La utilización de esta técnica de estimación alternativa requiere también que se evalúe el ajuste del modelo de varias formas.

La medida del ajuste global de cómo se ajusta el modelo viene dada por el valor de la verosimilitud (que es –2 veces el algoritmo del valor de verosimilitud y se representa por –2LL o –2 veces el logaritmo de la verosimilitud). Un modelo con buen ajuste tendrá un pequeño valor para –2LL, siendo cero el valor mínimo. Un ajuste perfecto del modelo tiene una verosimilitud de 1 y –2LL es cero. Los programas informáticos cuentan con contrastes automático para evaluar la significación de estas diferencias.

El contraste chi-cuadrado[3] para la reducción en el logaritmo del valor de verosimilitud proporciona una medida de mejora debido a la introducción de variable(s) independiente(s). Un modelo nulo, que es similar a calcular el total de la suma de los cuadrados utilizando sólo la media, proporciona el punto de partida para la comparación. Además de las contrastaciones estadísticas de los test de chi-cuadrado, se han construido varias medidas diferentes de tipo R^2 para representar el ajuste global del modelo, como lo hace el coeficiente de determinación de la regresión múltiple. El analista puede construir un valor "pseudo R^2" para la regresión logística similar al valor R^2 del análisis de regresión.

3 Estadístico Chi-cuadrado: Método de estandarización de datos en una *Tabla de Contingenica* comparando la frecuencia de la celda efectiva con la frecuencia de la celda esperada. La frecuencia de la celda esperada se basa en las probabilidades marginales de su fila y columna (probabilidad de una fila y columna entre todas las filas y las columnas).

El R^2 de un modelo logit se calcula como:

$$R^2_{logit} = \frac{-2LL_{nulo} - (-2LL_{model})}{-2LL_{nulo}}$$

Se puede evaluar el ajuste global de forma similar a la regresión múltiple y puede hacerse uso de varios métodos que utilizan la característica no métrica de la variable dependiente.

Un método es a través de los contrastes de clasificación (Hosmer y Lebenshow). Los casos se dividen en 10 clases aproximadamente iguales. Luego, el número de sucesos reales y predichos se compara en cada clase con el estadístico de chi-cuadrado . Este contraste proporciona una medida global de exactitud predictiva que no se basa en el valor de la verosimilitud, sino en la predicción real de la variable dependiente. El uso correcto de este contraste requiere un tamaño de muestra adecuado para asegurar que cada grupo cuenta al menos con cinco observaciones y nunca cae por debajo de uno. Además el estadístico chi-cuadrado es sensible al tamaño muestral, permitiendo por tanto, que esta medida encuentre diferencias estadísticamente muy pequeñas cuando el tamaño muestral crece

Contrastación de la significación de los coeficientes

La regresión logística puede contrastar la hipótesis de que un coeficiente sea distinto de cero como en la regresión múltiple. La regresión logística utiliza el estadístico de Ward que proporciona la significación estadística para cada coeficiente estimado de tal forma que se pueden contrastar hipótesis igual que en la regresión múltiple.

Semejanzas con la regresión múltiple

A pesar del hecho de utilizar una medida dependiente binaria y de que el resultado sea la predicción de pertenencia al grupo, el formato de la regresión logística es bastante parecido al de la regresión múltiple. Al igual que en la regresión, los datos categóricos y nominales pueden incluirse como variables independientes por medio de su codificación como variables ficticias.

CAPITULO III
APLICACIÓN DEL DATA MINING EN UN PROGRAMA DE FIDELIZACIÓN MULTIMARCA

Minería y Análisis de Datos aplicados en un Programa de Fidelización de clientes multimarca

El Programa de Fidelización Multimarca Travelpass

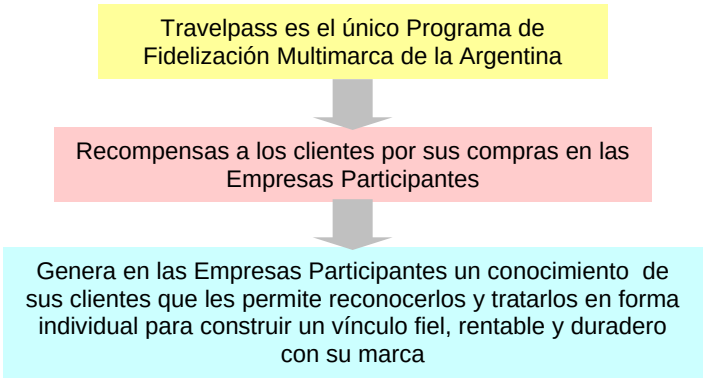
El Programa Travelpass es el primer y único programa de fidelización de clientes multimarca implementado en Argentina que - a través de una tarjeta cobranding - permite acumular puntos canjeables por recompensas.

El programa está sostenido por cuatro empresas socias fundadoras: Shell, Banco de Galicia, Supermercados Norte y el Grupo Telecom además de integrar a otras 11 empresas de diversos rubros: cadena de restaurantes, óptica, zapaterías, neumáticos, electrodomésticos, peluquería, central de reservas hoteleras, artículos de computación, jugueterías, farmacia. Los clientes Travelpass reciben puntos por sus compras en las empresas participantes que pueden canjear por recompensas del Programa.

La presentación tiene como propósito mostrar de qué modo las tecnologías de la información, los modelos estadísticos y de minería de datos conceptualmente desarrollados en la sección anterior y la investigación de mercado se integran en la realidad del negocio para optimizar la búsqueda de respuestas a las necesidades del área de marketing y control de gestión.

Presentación del Programa

El programa comenzó su actividad en septiembre de 1998 sostenido en primer lugar por cuatro empresas socias fundadoras: Shell, Supermercados Norte, Banco de Galicia y Grupo Telecom, este último integrando los servicios de telefonía residencial, celular e Internet. Además se sumaron otras empresas participantes de diferentes rubros.



Cómo surge el Programa

El Programa surgió como consecuencia de una inquietud de Shell que deseaba armar su propio esquema de fidelización. Ante la evidencia de Programas multimarca exitosos que había en el exterior, Shell vio una oportunidad en el mercado, logrando entusiasmar a otras empresas que inmediatamente se sumaron al proyecto. Es así como Supermercados Norte, Telecom. y Banco Galicia se sumaron al proyecto, además de otras Empresas Participantes.

La creación de Travelpass implicó dos años de desarrollo y la asignación de un presupuesto propio para publicidad, al que se añadieron las campañas que han preparado las Empresas Socias haciendo hincapié en el Programa.

La etapa de gestación incluyó también un **benchmarking internacional** que tomó en cuenta los principales programas del Reino Unido, Holanda, Canadá, Australia y Nueva Zelanda. De éste último país es justamente el software empleado para la administración del programa denominado Loyalty.

El objetivo del Programa es:

"retener a los clientes más valiosos y tratar de que concentren sus compras diarias en las empresas vinculadas a Travelpass. Además se busca **incrementar el cross traffic, motivando a los clientes a realizar sus compras en las empresas en las que no lo han hecho".**

Al hablar de Programas de Fidelización, cabe destacar que la principal diferencia entre Travelpass y los Programas existentes en el mercado argentino (DiscoPLus, ServiClub, JumboMás, y otros) consiste en la característica "**multimarca**". Este Programa, al tener muchas Empresas Participantes, cuenta con un atractivo especial ya que de esta forma los clientes poseen más oportunidades de sumar puntos rápidamente, concentrando sus compras habituales en esas compañías. Y este es uno de los puntos fuertes del Programa.

Visión:

Ser el programa líder en fidelizar el consumo de las familias al conjunto de Empresas Participantes del Programa.

Misión:

Lograr que los mejores clientes de Travelpass, motivados por las recompensas y el reconocimiento, concentren sus necesidades de compras y servicios en las Empresas Participantes del Programa, de manera que las mismas mejoren cuantitativamente sus negocios en forma sostenida y medible.

Estrategia General:

Generar en las Empresas Participantes un conocimiento tal de los clientes que les permita recompensarlos y tratarlos en forma individual para construir un vínculo sólido, rentable y duradero con su marca.

El funcionamiento del programa es sencillo: los clientes acumulan puntos en cada una de las empresas participantes, los cuales son canjeables por recompensas en el Programa. La posibilidad de sumar en una amplia variedad de empresas permite una acumulación más rápida de puntos y la obtención de recompensas en menos tiempo.

Recompensas para el tiempo libre

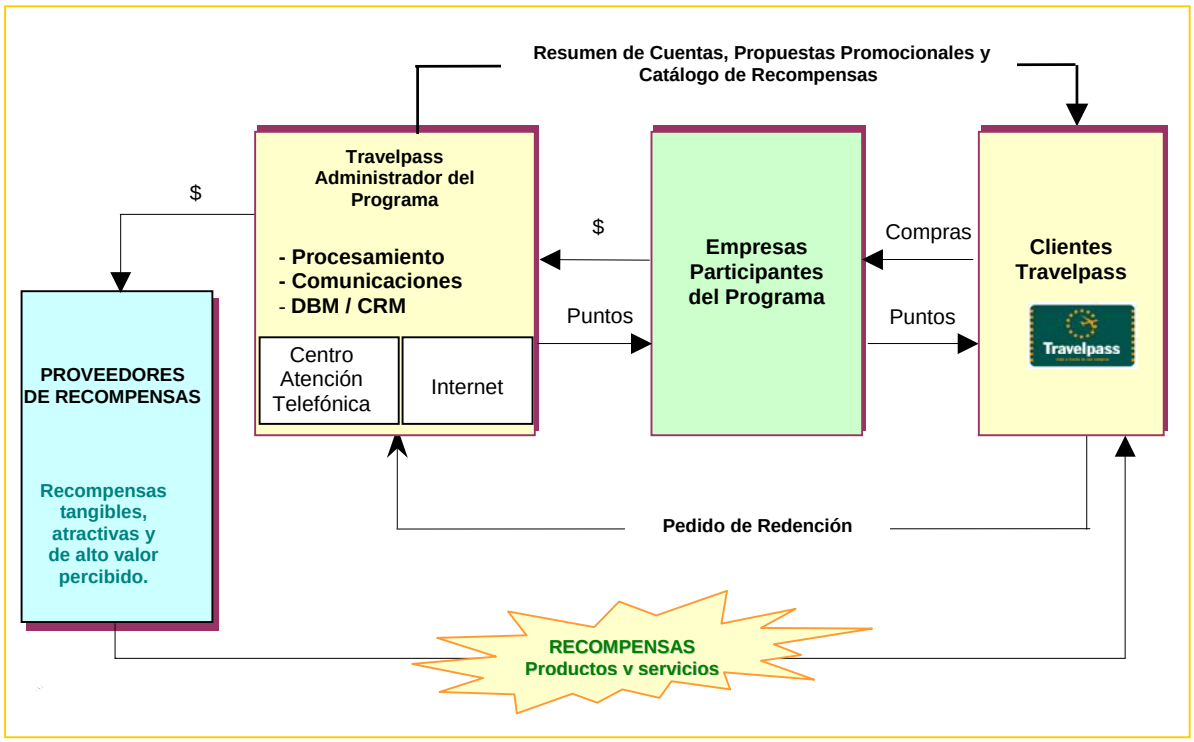
El programa nació con la idea de dar recompensas relacionadas con viajes y estadías. Posteriormente se fueron incorporando otras empresas vinculadas al concepto de "disfrute del tiempo libre": CD's, cenas, bicicletas. Son recompensas más tangibles y cercanas. A lo largo del Programa se fueron incorporando también electrodomésticos, calzados y artículos para camping que rápidamente se transformaron en los más canjeados.

Las Empresas Participantes

Las empresas mantienen su propia relación de pesos por puntos, dependiendo de su perfil comercial (ticket promedio, facturación y rentabilidad) y sus objetivos de fidelización en el Programa Travelpass.

SHELL $ 10 = 5 Puntos	**SUPERMERCADOS NORTE** $ 20 = 10 Puntos	**TELECOM** $ 20 = 5 Puntos	**TELECOM PERSONAL** $ 20 = 10 Puntos	**TELECOM ARNET** $ 10 = 5 Puntos
BRIGESTONE FIRESTONE $ 10 = 10 Puntos	**GRIMOLDI** $ 20 = 10 Puntos	**VENTURA** $ 50 = 25 Puntos	**LLONGUERAS** $ 20 = 20 Puntos	**LA CABALLERIZA** $ 10 = 10 Puntos
LA OPTICA EXPRESS $ 10 = 10 Puntos	**EL MUNDO DEL JUGUETE** $ 20 = 10 Puntos	**COMPUEXPERT** $ 50 = 25 Puntos	**CLICK HOTELES** $ 10 = 10 Puntos	**HIPERCENTRO DEVOTO** 1 Hipermilla = 1 Punto

Estructura Funcional del Programa



Los Clientes Travelpass

- Los clientes concentran sus compras en las Empresas Participantes, presentan su Tarjeta Travelpass y por ello reciben Puntos Travelpass.

Las Empresas Participantes

- Las Empresas Participantes envían a Travelpass los Puntos entregados a sus clientes.
- Desarrollan sus Estrategias de Relacionamiento con sus propios Clientes
- Genera acciones de Marketing Directo, a través de la base de clientes Travelpass.

Responsabilidades de Travelpass como Ente Administrador

- **Procesamiento de Puntos**
  - Travelpass procesa los Puntos en las Cuentas de los Clientes y factura los Puntos enviados por las Empresas Participantes.
  - Administra el procesamiento y almacenamiento de todas las transacciones.

- **Comunicaciones**
  - Desarrolla y mantiene los canales de comunicación con el cliente:
    - Centro de Atención Telefónica
    - Sitio web.

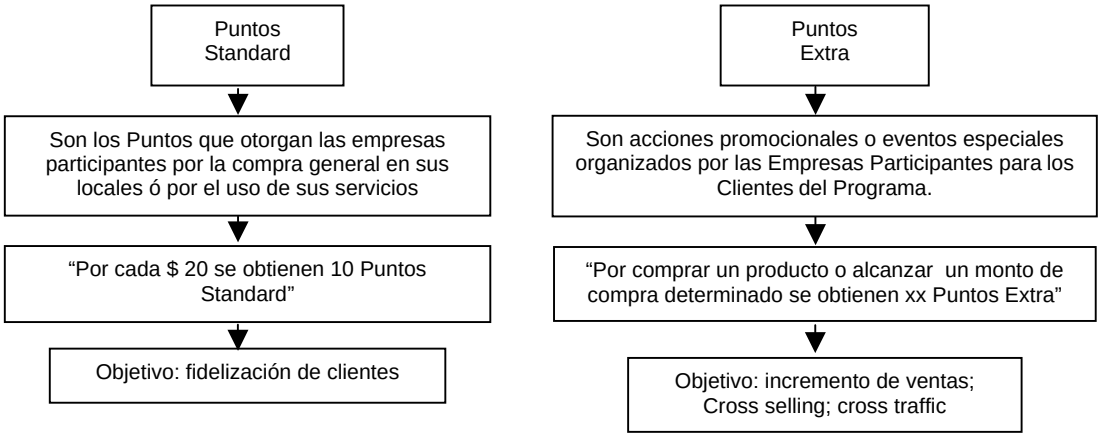
- **DBM - CRM**
  - Desarrolla la Estrategia General de Relacionamiento con los Clientes Travelpass.
  - Desarrolla acciones de DBM.

- Brinda soporte de marketing directo y comercial a las Empresas Participantes para el desarrollo sus propias estrategias de fidelización y relacionamiento con sus clientes.

- **Gestión de Recompensas**
  - Definición de la Estrategia Global de Recompensas del Programa.
  - Administración de la Comunicación de las Recompensas.
  - Negociación con Proveedores para la compra de Recompensas.

Puntos Standard y Puntos Extra.

Las Empresas Participantes entregan Puntos Standard y Puntos Extra:

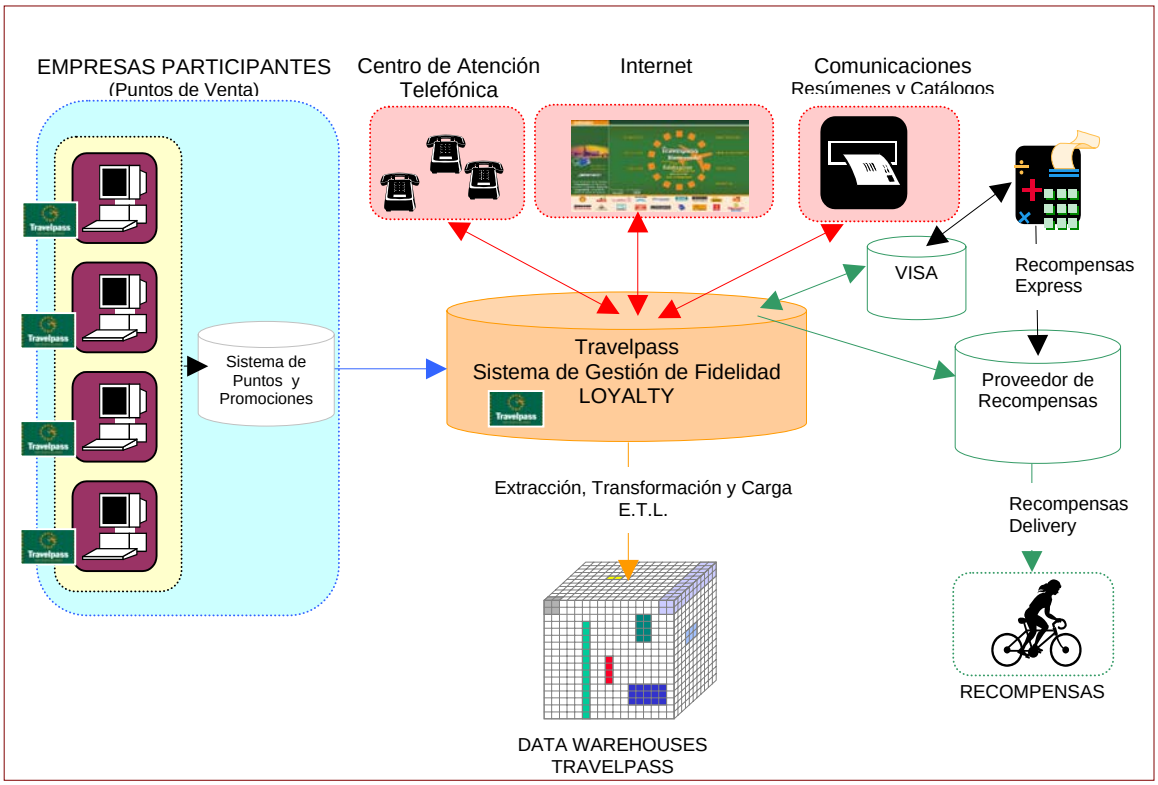


Los puntos acumulados tienen una vigencia de 24 meses, a partir de la fecha en que han sido acreditados.

Estructura Operacional del Programa.

Desde sus comienzos Travelpass comenzó a desarrollar el área de análisis de datos como soporte de las actividades de database marketing, CRM y control de gestión. Estas actividades sirvieron además de soporte para el diseño e implantación del Datawarehouse que facilita el acceso a los datos.

ESTRUCTURA FUNCIONAL DEL PROGRAMA



- **Puntos de Venta**
  El cliente presenta su tarjeta Travelpass en los Puntos de Venta adheridos de las empresas participantes del Programa. El número de Tarjeta Travelpass se asocia a la transacción de compra realizada por el cliente.

- **Sistema de Puntos y Promociones**
  Es el sistema que toma las transacciones de los Puntos de Venta y asigna los Puntos Standard y los Puntos Extra, según los parámetros de promoción previamente definidos.
  Genera la información en el formato requerido por el Sistema Transaccional de Fidelidad Loyalty.

- **Sistema de Gestión de Fidelidad (LOYALTY)**
  Es la Base de Datos Relacional que administra y almacena todas las transacciones realizadas por el cliente, sea por acumulación de puntos (créditos); transacciones por canjes de recompensas (débitos) y otras.
  Almacena la historia transaccional de cada uno de los clientes del Programa desde su fecha de activación.

- **Almacenamiento de Datos - Datawarehouse**
  El objetivo principal de la creación del Datawarehouse en Travelpass fue la integración de los datos de toda la empresa, en un formato asequible para el análisis de información.

El rol fundamental del Data Warehouse es proveer datos para soportar la toma de decisiones:

- ✓ Ha expandido el impacto del dato mediante el foco en el alcance, precisión y accesibilidad del mismo.
- ✓ Mayor horizonte temporal.
- ✓ Facilita otras aplicaciones como OLAP basada en BD Relacionales o Multidimensionales.
- ✓ Optimizado para contestar consultas complejas.
- ✓ Facilita el concepto de Data Mining.

Características del Datawarehouse de Travelpass:

- Posee una vista multidimensional.
- Es transparente al usuario.
- Accesible.
- Reportes consistentes.
- Arquitectura Cliente /Server-Web.
- Dimensionalidad Genérica.
- Manejo dinámico de matrices dispersas.
- Soporte multiusuario.
- Operaciones a través de varias dimensiones.
- Manipulación intuitiva de los datos.
- Reportes flexibles.
- Dimensiones ilimitadas, agregaciones

- **Sistema de Recompensas**

El canje de recompensas se realiza bajo tres modalidades:

1. **Modalidad Express**
2. **Delivery**
3. **Canjes a través de la web**

1. Modalidad Express

La modalidad express es la modalidad de **canje inmediato**, que se ha estructurado sobre la red de Pos de Puntos de Venta –POS-
El Centro autorizador es Visa Argentina y se han homologado los terminales de Puntos de Venta de Mastercard, AMEX y Visa.

Etapa Tecnológica	Proceso	Recursos Tecnológicos
Canje Express	**Integración con la red de POS de Visa** **Homologación de los Terminales POS de AMEX, Mastercard y VISA**	**POS** **Centro Autorizador** **Call Center** **Autorizaciones**

El cliente presenta la Tarjeta Travelpass en los locales adheridos a la recompensa express y previa consulta de los Puntos Travelpass disponible en su Cuenta, se procede a la opción "Canje". Mediante una comunicación on-line con el Centro Autorizador de Visa, se descuentan los Puntos automáticamente de la Cuenta y el cliente puede gozar de su recompensa al instante.

2. Modalidad Delivery

El cliente Travelpass solicita el envío de la recompensa a través del Centro de Atención Telefónica de Travelpass.

3. Canjes a través de la web

El cliente ingresa a www.travelpass.com.ar y realiza su canje on line.
Sólo pueden canjear vía Internet los clientes registrados en la web, es decir, que hayan recibido su Clave de Identificación Personal.

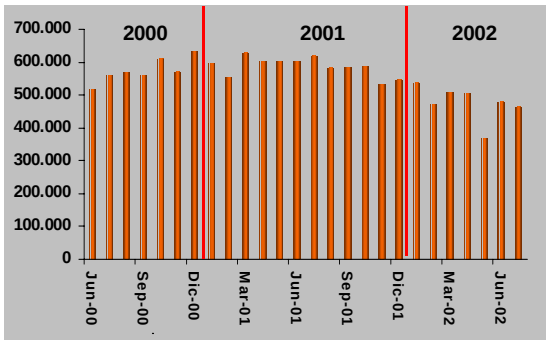
Indicadores del Programa: Su Evolución

El programa tiene una cobertura nacional y hasta Agosto de 2002 su Data Warehouse almacenaba información relativa a:

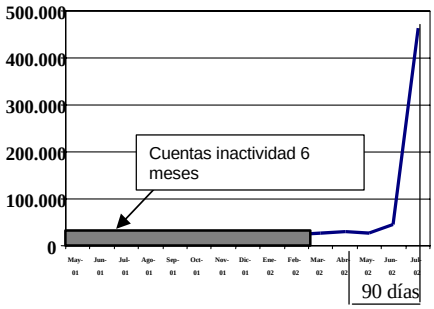
- ➔ **1.225.000 cuentas activas en el programa.**
- ➔ **2.950.000.000 millones de puntos emitidos**.
- ➔ **74.000.000 millones de transacciones** procesadas.
- ➔ **4.220.000.000 millones de pesos** por compras de clientes.
- ➔ **2.800** puntos de venta con identificación Travelpass.
- ➔ **550** Centros de Canje Express.
- ➔ **500.000 recompensas canjeadas**
- ➔ **38%** de índice de redención de Puntos.

Evolución de los Principales Indicadores

Cuentas con Compras en el Mes

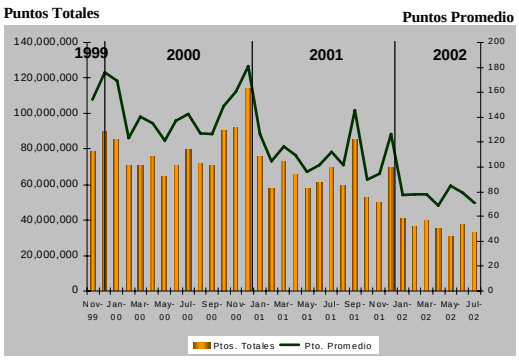


Recencia



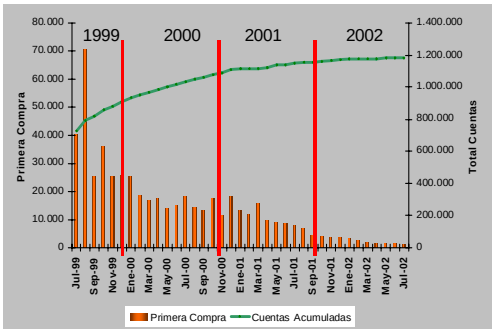
535.823 cuentas usaron Travelpass en los últimos 3 meses.(Recency)

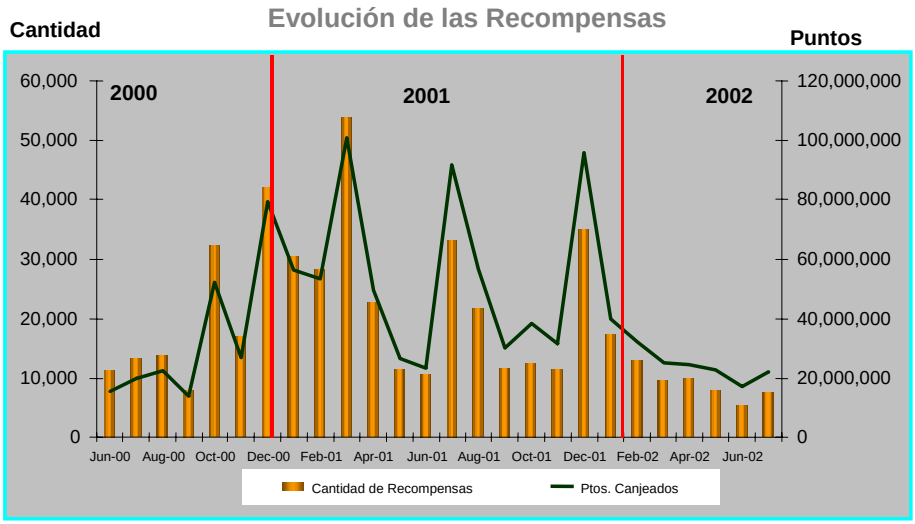
Puntos Totales y Punto Promedio por Cuenta



Primera Compra de Clientes

Activación Enero-Julio 2002 14.147 cuentas





Índice de Redención de Puntos (Ago'02): **38%**

Principales indicadores de Performance del Programa:

- **Cuentas con compras en el mes:** mide la cantidad de cuentas que presentaron su tarjeta en las compras del mes. Es un indicador de la evolución mensual del programa.

- **Recencia**: es un indicador de la cantidad de cuentas que realizaron su última compra en el período más reciente. Como el programa concentra un amplio espectro de Servicios y retailers, con negocios de frecuencias variadas, se convino en medir una recencia de 90 días a nivel general del programa. En consecuencia, este indicador marca la cantidad de cuentas que realizaron compras en el último trimestre. Las cuentas que no realizaron compras en los últimos 6 meses son cerradas por inactividad.

- **Primera compra de Clientes**: es un indicador de la cantidad de cuentas que se activan (dan de alta) en el programa mes a mes. Como puede observarse en el gráfico, al inicio del proyecto la activación de cuentas era muy intensa como respuesta a una fuerte campaña de lanzamiento y adhesión de clientes.

- **Evolución del canje de recompensas:** el canje de recompensas es uno de los indicadores de gestión más importantes del Programa.- El control del canje de recompensas por sobre las demás variables de gestión (restantes costos fijos y variables) mantiene el equilibrio financiero del programa y asegura un nivel de contingencia aceptable.

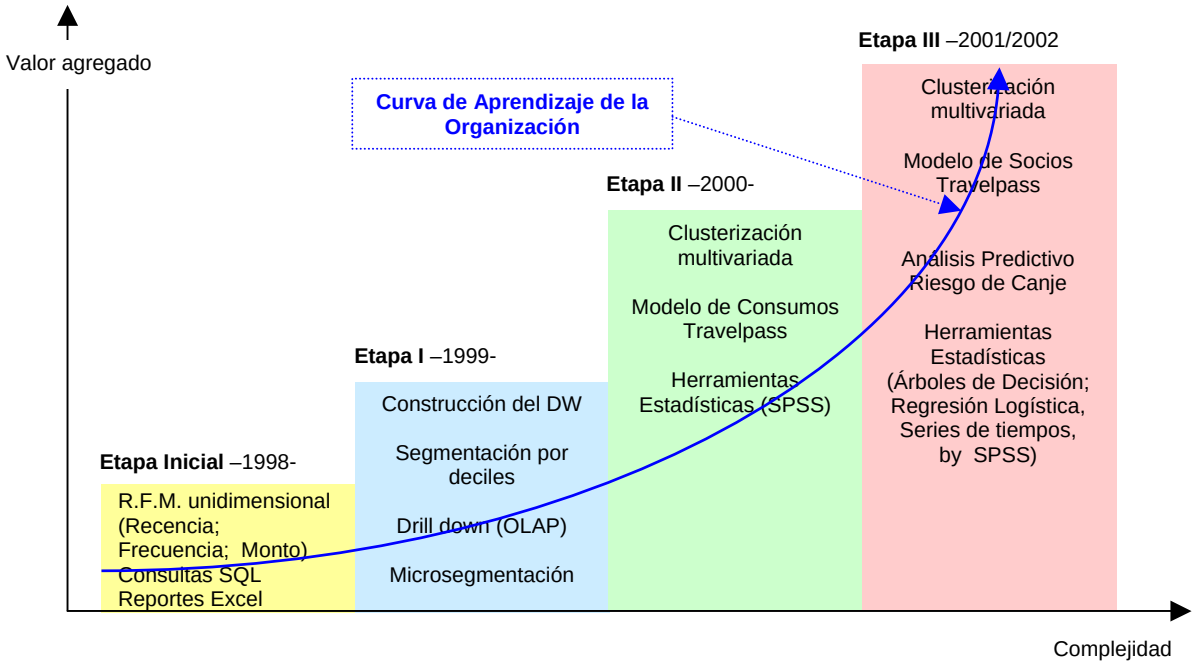
  Los "picos" que pueden observarse son consecuencia de la comunicación enviada a los Clientes Travelpass. Cada envío de Marketing Directo es un disparador inmediato del canje de recompensas.

  El Índice de redención indica la relación existente entre los puntos canjeados y los puntos totales acumulados en el Programa.

  A lo largo del proyecto se verá de qué manera los modelos de análisis de datos permitieron controlar estas variables de gestión.

Evolución del análisis de datos en el Programa Travelpass

El área de análisis de datos evolucionó desde sus comienzos en los que solamente se producían reportes simples obtenidos mediante tablas dinámicas Excel y SQL hasta la actualidad en la que se desarrollan modelos descriptivos y predictivos a partir de nuevas tecnologías de Data Mining.



Cuadro Síntesis de la Evolución del Análisis de Datos:

Año / Etapa	Procesos principales	Recursos Tecnológicos
1998- Etapa Inicial **Generación de Datos**	Altas de Clientes Procesamiento de Puntos Resúmenes de cuenta Emisión de certificados de canje Producción de Tarjetas R.F.M. Reportes Excel y Consultas SQL	Bases Relacionales de alta capacidad Servidores de Alta disponibilidad Enlaces de Comunicaciones Puntos de Captura de transacciones de Venta Producción y embozado de tarjetas / card carriers
1999- Etapa I **Segmentación Univariada**	Clasificación del consumo por deciles (segmentación univariable) Acciones Segmentadas Data-cleaning	Datawarehouse (Base Sybase) Herramientas OLAP (Brio)
2000-Etapa II **Segmentación Multivariada**	Generación de Variables RFM Modelo de Consumos Travelpass Contactos Segmentados	Datawarehousing Uso de herramientas estadísticas SPSS Principio de Data mining
2001/2002- Etapa III **Modelo de Segmentación Global y Modelo de Socios**	Clusterización multivariada Atención segmentada	Data Mining (SPSS)

Proyectos de Data Mining en el Programa Travelpass

A principios del año 2000 se inicia el primer proyecto de Data Mining que introduce a Travelpass en la etapa de exploración y análisis de los datos para encontrar patrones y reglas de negocio. El Análisis de Perfiles de Clientes fue el proceso que dio origen a las actividades de Business Intelligence como soporte a las áreas de Marketing y Control de Gestión del Programa.

Se desarrolló un Modelo Descriptivo Multivariado orientado a la SEGMENTACIÓN DE CLIENTES TRAVELPASS, a partir de las variables que describían su comportamiento de consumo.

La segmentación de los clientes: un modelo de Clustering Multivariado

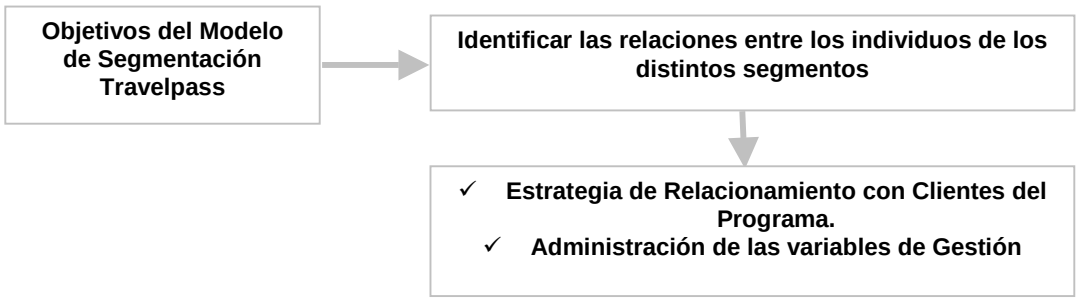
Como reseña de la Evolución del Análisis de Datos en Travelpass puede observarse que la segmentación de los clientes ha ido madurando sobre la base de principios metodológicos aceptados:

El enfoque de Segmentación Multivariada se denomina “data driven segmentation”, es decir, los datos mandan sin recibir supuestos ni restricciones a priori y es el proceso que se describirá a continuación.

OBJETIVOS DE LA SEGMENTACIÓN DE CLIENTES TRAVELPASS

Conceptualmente, el objetivo fundamental del análisis cluster es la obtención de un conjunto de objetos en dos o más segmentos, basándose en su similitud para un conjunto de características especificadas.

En el caso particular del Modelo de Consumos Travelpass (nombre que le fue posteriormente conferido), el objetivo inicial de la segmentación era identificar aquellos clientes de mayor fidelidad con el programa –clientes de mayor valor-, para describir luego su perfil sociodemográfico y establecer la Estrategia de Relacionamiento con los Clientes en el marco de Control de Gestión del Programa.-



SELECCIÓN DE LAS VARIABLES PARA EL ANÁLISIS CLUSTER

La selección de las variables utilizadas para caracterizar los objetos / sujetos a agrupar está directamente vinculada al objetivo del análisis cluster. Tanto en sentido confirmatorio como exploratorio, el resultado del modelo queda restringido a las variables utilizadas en el proceso. Los segmentos derivados reflejan la estructura inherente de los datos sólo como definida por las variables seleccionadas.

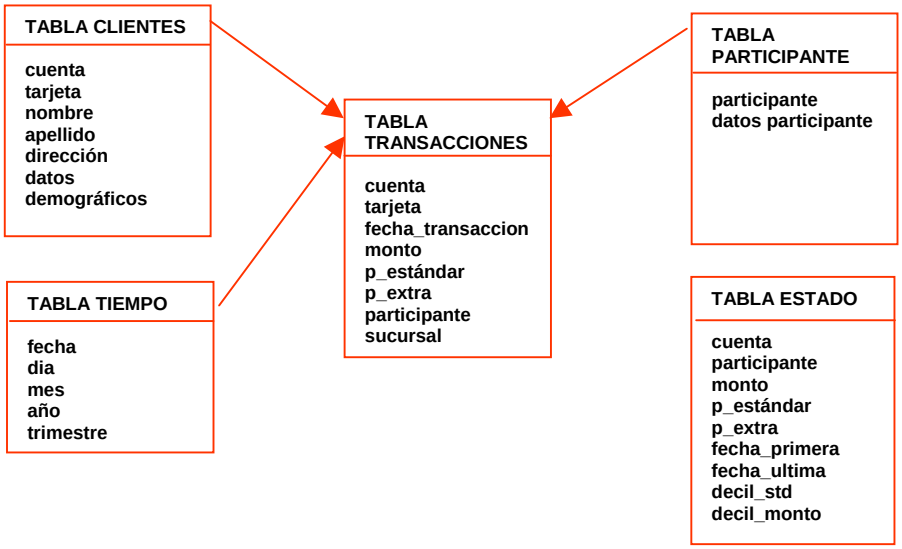
Se trataba de un problema de **Clustering** o aprendizaje "no supervisado" ya que se pretendía agrupar a los clientes según los indicadores de consumo que mejor los describían.

- **R (recency)**
- **F (frecuency)**
- **M (Monto gastado en el Programa: Puntos por mes)**
- **IM (Índice Multimarca)**

El ejercicio se planteó sin conocer de antemano las "reglas" de formación de los clusters, utilizando entonces una metodología "data driven" que no impusiera a los datos condiciones "a priori".

IDENTIFICACIÓN DE LAS FUENTES DE DATOS

El proceso de data mining requiere de datos. Como el Data Warehouse de Travelpass aún estaba en etapa de diseño, se partió de la Base de Datos Relacional del Sistema de Fidelidad LOYALTY (Sistema Operacional). Los datos provenían de las siguientes Tablas de datos:



OBTENCIÓN, LIMPIEZA Y TRANSFORMACIÓN DE LOS DATOS

Tan pronto como se tuvo acceso a los datos se analizó el perfil de los mismos para determinar su grado de consistencia y homogeneidad.

La primera etapa de preparación y data cleaning de los datos insumió una porción importante de recursos humanos y tiempo, pero como existía plena conciencia de que los resultados del proceso dependían de la calidad de los datos, resultaba necesario asegurar su consistencia antes de la aplicación del algoritmo de clustering.

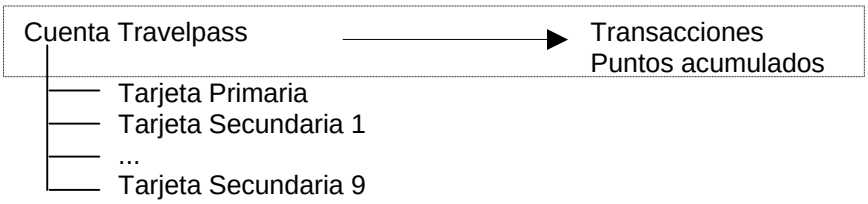
En la primera etapa se llevaron a cabo las siguientes tareas de **Data Cleaning**:

- ✓ Agrupar transacciones de la misma cuenta que tenían el mismo participante -día-hora.
- ✓ Filtrar transacciones inválidas (Puntos Standard = 0 y Puntos Extra = 0).
- ✓ Filtrar transacciones con fechas incorrectas.

El nivel de granulidad de los datos

El nivel de granulidad de los datos se refiere a la definición del tamaño de la unidad de análisis para data mining. Implica identificar cuál es la unidad "cliente" en el proceso de análisis.

Desde el punto de vista estructural, el Programa Travelpass está diseñado para que toda la familia pueda sumar puntos en la misma Cuenta Travelpass. Una Cuenta Travelpass involucra una Tarjeta Primaria y puede además contener hasta 9 Tarjetas Secundarias. El resumen de transacciones realizada por cualquiera de las Tarjetas Travelpass se realiza a nivel de la Cuenta Travelpass a la que están vinculadas, así como los Puntos Travelpass sumados en cualquiera de las empresas participantes del Programa.



En virtud de ello, la unidad de estudio considerada en el análisis es la Cuenta Travelpass, como entidad que representa al Cliente Travelpass (familia)

SUMARIZACIÓN DE LOS DATOS Y AGREGADO DE VARIABLES DERIVADAS

El nivel apropiado para la sumarización de los datos está directamente relacionado con el nivel requerido para el análisis.

En este caso se realizaron las siguientes sumarizaciones de datos por Cuenta Travelpass:

- **Promedio de días entre transacciones (frecuencia).**
- **Fecha primera y última transacción de la cuenta.**
- **Ticket promedio de todas las transacciones que lo informan.**
- **Puntos Standard totales de cada cuenta.**
- **Multimarca:** Índice que indica cantidad de empresas participantes en donde la cuenta realizó transacciones**.**

Además se obtuvieron las siguientes **variables derivadas**:

- **Recency**: Es la diferencia en días entre la Fecha del análisis y la Fecha de la última transacción de la cuenta.
- **Puntos por mes:** Es el cociente entre Puntos Standard totales y meses de permanencia en el programa. (Nótese que en el análisis se involucran los Puntos Standard como referente del nivel de fidelidad de los clientes en el Programa.
- **Activación:** días transcurridos desde la primera transacción en el Programa, al momento del estudio.

Finalmente, en la siguiente Tabla se observa el formato de la tabla de datos sometida al clustering.

Tabla Final de Datos para la Segmentación

Cuenta	Ticket prom.	Frecuencia	Recency	Std / mes	Multimarca
111111	24,5	152,88	34,28	3,92	3
111112	299,7	missing	601,74	140	1
111113	293,35	54,55	89,87	46,35	1
111114	16,69	28,5	515,74	13,16	1
111115	2887,45	40,57	110,18	1066,7	2
111116	87,73	2,46	11,25	515	1
111117	10	missing	129,79	5	1
111118	19,76	12,04	5,23	24,29	1
111119	27,31	1,38	277,43	195	1
111120	52,74	4,19	19,91	186,36	5
111121	46,14	11,23	16,84	58,4	3
111122	61,92	4,54	7,24	143,94	10
111123	40	9,75	351,83	80	1
111124	72,57	18,53	41,58	36,77	7
111125	46,67	21,94	390,86	47,86	1
111126	628,47	40,15	15,94	351,18	1
111127	37,59	1,5	12,14	377,46	1
111128	32,4	6,07	7,98	78,73	1
111129	32,1	9,68	7	49,92	3
111130	22,64	29,65	44,83	15,18	1
111131	41,88	6,83	90,24	100,45	1
111132	missing	missing	55,13	0	1

♦ **DETECCIÓN DE DATOS AUSENTES Y CASOS ATÍPICOS**

La exploración univariada inicial de los datos permitió la identificación de **Datos Ausentes** (missings) y **Casos Atípicos** (outliers) para los que se adoptó el siguiente criterio:

1. Datos Ausentes (missings)

Los datos ausentes son habituales en el análisis de múltiples variables. El principal desafío es enfrentarse a los resultados que ellos producen y la principal ocupación es determinar las razones que subyacen en ellos. Esta necesidad sugiere entender el proceso de ausencia de los datos para seleccionar el curso de acción más conveniente.

En este caso se identificaron "**datos ausentes**" presentes en la variable "**Frecuencia**" para las cuentas con **1 sola transacción** (Recuérdese que "Frecuencia" es igual al promedio de días entre transacciones).

Para determinar las razones que subyacen en las Cuentas con 1 sola transacción se siguieron dos líneas de pensamiento:

a) Se trataba de **Cuentas Nuevas**

b) Se trataba de **Cuenta con una baja participación en el Programa**.

El paso siguiente fue analizar la **Fecha de Activación de la Cuenta**, es decir, los días transcurridos desde la primera transacción en el Programa, al momento del estudio. Se convino asumir 120 días como métrica razonable para evaluar el comportamiento de la Cuenta en términos de transacciones procesadas.

De la combinación de ambas variables "**Frecuencia**" y "**Fecha de Activación**" surgieron dos segmentos **pre-definidos** que, una vez determinados se **excluyeron** del análisis de segmentación general:

a) Cuentas con **1 sola transacción** y **menos de 120 días de Activación en el Programa** se incluyeron en un segmento ó cluster que se denominó **"New".**
Estas cuentas representaron el **2% del total** de cuentas sujeto a análisis.

b) Cuentas con **1 sola transacción** y **120 días ó más de Activación en el Programa** se consideraron Cuentas con baja participación en el Programa y se los incluyó en un segmento ó cluster que se denominó **"Descarte".**
Estas cuentas representaron el **14% del total** de cuentas sujeto a análisis.

Los totales señalados en el párrafo anterior se muestran en la siguiente tabla de porcentajes:

Días de Activación (días transcurridos desde la primera transacción en el programa)	**% Cuentas con más de 1 transacción**	**% Cuentas con 1 transacción**	**Total**
120 días o más	→ 75%	14%	89%
Menos de 120 días	9%	2%	11%
	84%	**16%**	**100%**

Luego de este tratamiento para los "casos ausentes", quedó una base conteniendo el 84% del total de las **Cuentas Travelpass** inicialmente consideradas.

2. **Casos Atípicos** (Outliers)

Los "**casos atípicos**" son observaciones de una combinación única de características identificables que les diferencia claramente de las otras observaciones. Los casos atípicos deben ser analizados en el contexto del estudio y evaluados en función de la información que proporcionan, antes de caracterizarlos como benéficos o problemáticos.

En estos casos, es ocupación del analista de datos determinar las razones que subyacen en los "casos atípicos" y seleccionar el curso de acción más conveniente.

En el caso de estudio se identificaron "**casos atípicos"** presentes en las variables **Recency** y **Puntos por mes**.

Se discretizó a las variables en cuartiles y se asimilaron los outliers a los extremos según se indica a continuación:

Variable / Código	**1**	**2**	**3**	**4**	**5**
Recency Referencia **R**	Hasta 30 días	30 a 90 días	90 a 120 días	Más de 120 días	
Frecuencia Referencia **F**	Hasta 7 días	7 a 14 días	14 a 30 días	Mas de 30 días	
Puntos por mes Referencia **P**	100 o más	50 a 100	20 a 50	Hasta 20	
Índice Multimarca Referencia Mu	1	2	3	4	5 ó más

Selección de las variables para el análisis cluster

A partir de estas transformaciones se obtiene entonces un conjunto de indicadores ordinales para cada Cuenta Travelpass. Así por ejemplo un cliente valorizado por las siguientes categorías está calificado en los mejores puntajes para cada variable:

R1, F1, Mu5+ y P1

Diseño de la investigación mediante el análisis cluster

El análisis cluster agrupa a los individuos y a los objetos en segmentos, de tal forma que los objetos del mismo segmento son más parecidos entre sí que a los objetos de otro segmento. El análisis cluster intenta maximizar la homogeneidad de los objetos ó sujetos dentro de los segmentos y a la vez maximiza la heterogeneidad entre los agregados.

Obtención de segmentos y valoración del ajuste conjunto

Con las variables seleccionadas y la matriz de similitud entre las variables calculadas, se aplicó el Algoritmo de Clustering residente en el software estadístico para data mining provisto por SPSS.

El criterio esencial de todos los algoritmos es maximizar las diferencias entre los segmentos, relativa a la variación dentro de los mismos.

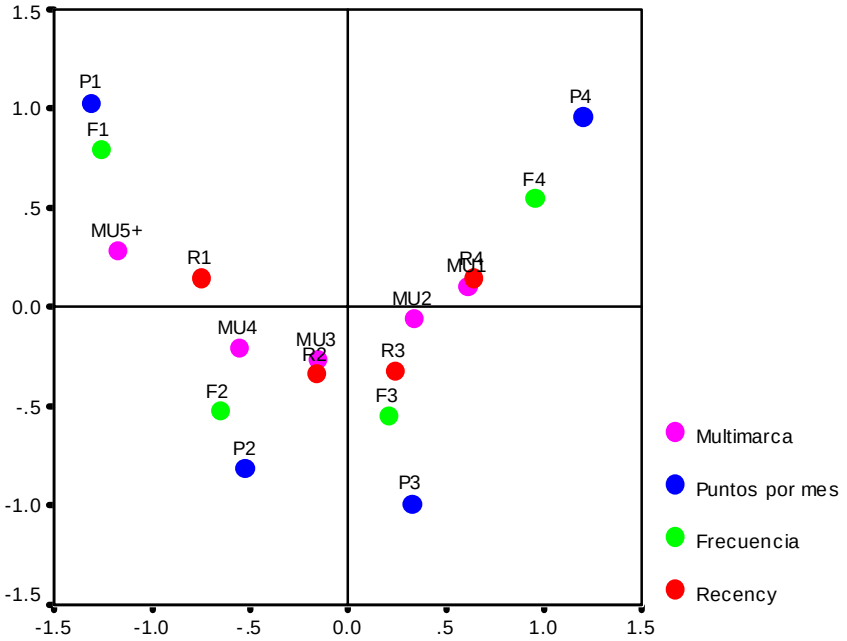
El Algoritmo de Clustering aplicado al caso, es un Algoritmo de tipo **No Jerárquico** frecuentemente denominado como aglomeración de **K-MEDIAS**. Este tipo de algoritmos asigna los objetos una vez que el número de segmentos está especificado. Inicialmente se planteó la posibilidad de trabajar con 6 segmentos pero luego de varias “corridas” del modelo se optó por migrar a una solución de 4 segmentos.

Dada la cantidad de clusters o segmentos a formar, el primer paso desarrollado por el algoritmo fue seleccionar una semilla de segmento como centro del segmento inicial, y todos los objetos (individuos) dentro de una distancia umbral previamente especificada se incluyen dentro del segmento resultante. El algoritmo entonces selecciona otra semilla de segmento y la asignación continuó hasta que todos los objetos estuvieran debidamente asignados. Los objetos pueden entonces asignarse si están cercanos a otro segmento que no sea el original.

Para aplicar el algoritmo de clustering disponible en la herramienta de análisis (K-Means) resultó necesario definir una métrica de distancia euclideana entre los individuos (clientes). Para ello se trataron los puntajes ordinales con un método de Componentes Principales Categóricas (CATREG) que permite obtener las coordenadas en un plano factorial de las categorías de variable y de los individuos como se muestra en los Mapas Perceptuales[3]1 y 2.

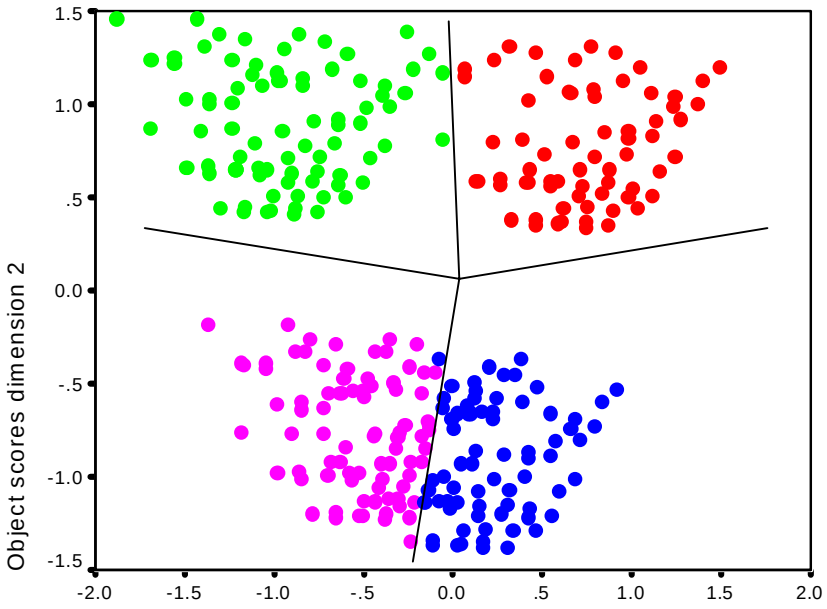
[3] Desde su aparición en el campo del marketing , los "mapas perceptuales" se han convertido en uno de los instrumentos de comunicación de análisis estadísticos más extensamente utilizados para proporcionar información dirigida al proceso de toma de decisiones. Entre numerosas aplicaciones, un mapa perceptual es extremadamente útil para disponer de una amplia visión de los puntos fuertes u ocupación de los segmentos de clientes. Un "mapa perceptual" se entiende como la representación gráfica de los resultados estadístico-matemáticos que permiten modelizar percepciones.

Mapa Perceptual 1: PLANO COORDENADO DE CATEGORÍAS DE VARIABLES COMPONENTES PRINICIPALES CATEGÓRICAS (CATREG)



Variable Principal normalization.

Mapa Perceptual 2: SEGMENTOS MAPA DE INDIVIDUOS



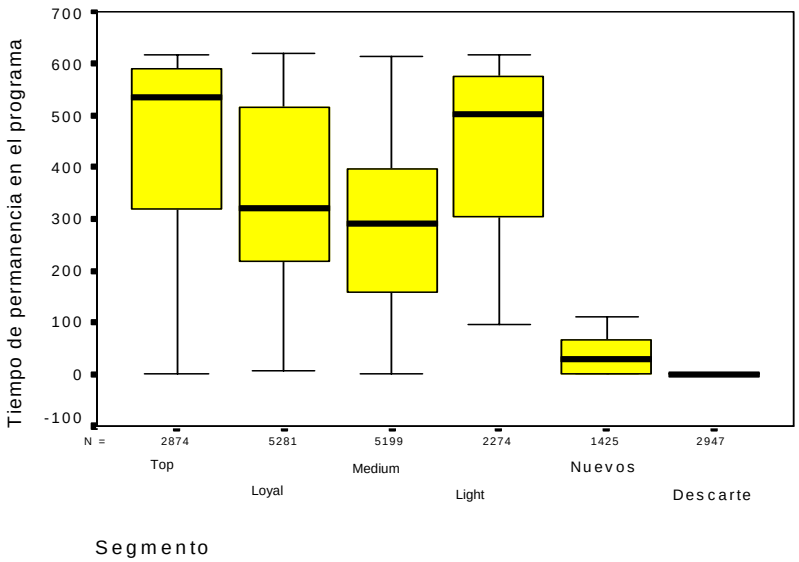
Para el **PLANO COORDENADO DE CATEGORÍAS DE VARIABLES** se consideró a las variables medidas en una escala múltiple nominal (sin ordinalidad).

En el mapa de categorías se observa que la primera dimensión - que explica el 61% de la variabilidad total – refleja la ordinalidad de las categorías de todas las variables a pesar de que no fue un supuesto de análisis.

El mapa de individuos muestra la separación de los clientes en 4 clusters.

Un algoritmo de K-Medias que se aplica entonces sobre las dimensiones resultantes del análisis en componentes principales conduce a 4 segmentos que posteriormente se denominaron: Top, Loyal, Medium, Light.

Análisis por segmento del Tiempo de Permanencia en el Programa



Otra de las tareas que se ha llevado a cabo es entender el carácter y la diferencia entre dos o más segmentos para la variable Tiempo de Permanencia en el Programa.

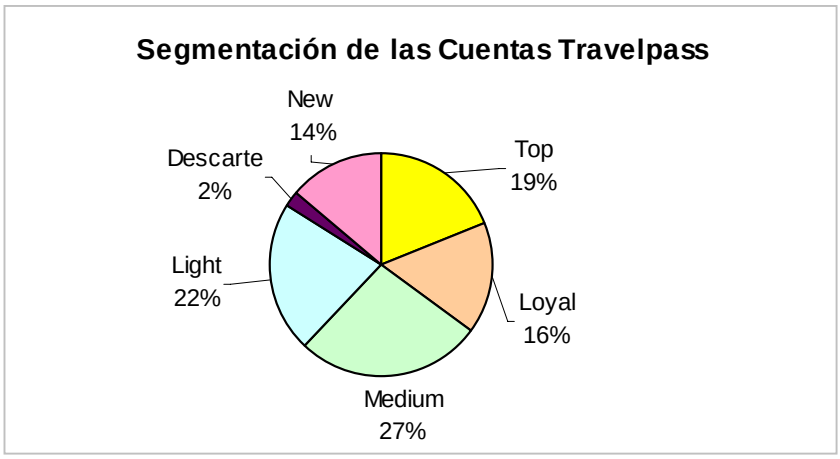
En este caso se necesitaba entender cómo se distribuían los valores para cada grupo y si existían suficientes diferencias entre ellos como para tener significación estadística.

El método que se utiliza para ésta tarea es el **gráfico de cajas** (boxplot), una representación gráfica de la distribución de los datos. Los límites superior e inferior de la caja marcan los cuartiles superior e inferior de la distribución de datos. Por tanto, la longitud de la caja es la distancia entre el primer y el tercer cuartil, de forma que la caja contiene el 50% de los datos centrales de la distribución. La línea dentro de la caja señala la posición de la mediana. En el gráfico puede observarse que la mediana de los clientes Top tiene el tiempo de Permanencia en el programa, con una asimetría hacia los valores superiores de la variable.

El tamaño de la caja indica la extensión de las observaciones. Las líneas que se extienden desde la caja (llamadas bigotes) representan la distancia entre la mayor y la menor de las observaciones que están a menos de un cuartil de la caja. Los casos atípicos son observaciones que se sitúan entre 1.0 y 1.5 cuartiles fuera de la caja.
Los valores extremos son aquellas observaciones mayores que están a 1.5 cuartiles fuera de los límites de la caja.

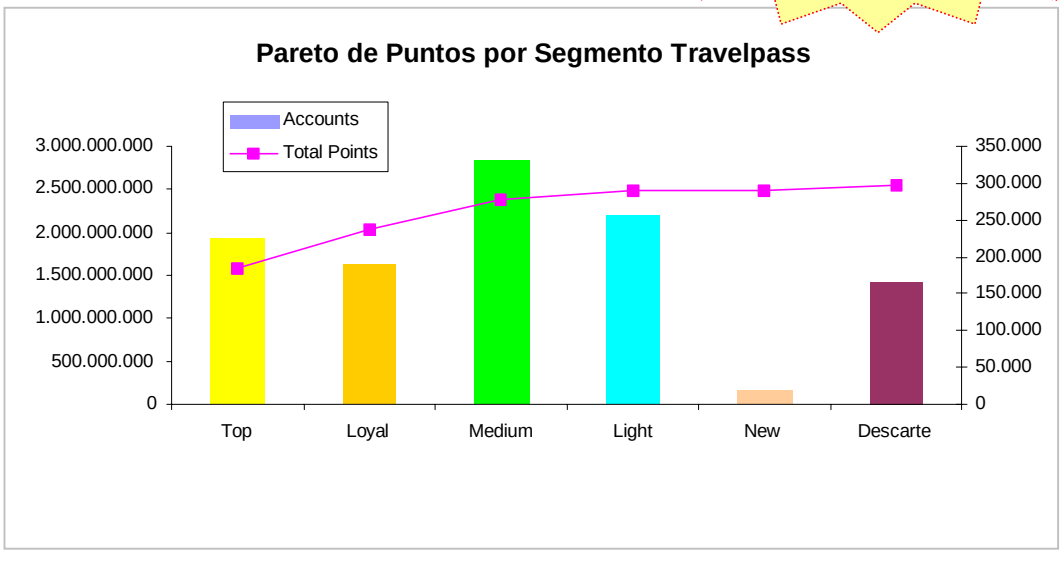
Interpretación de los Segmentos

Para confirmar la claridad de los cluster obtenidos se ha representado el escenario de cuentas y puntos en un diagrama de Pareto. En el diagrama de Pareto puede observarse la participación de estos segmentos en el total de puntos entregados por Travelpass. El 80% de los puntos emitidos se concentra en 2 segmentos (Top y Loyal) que representan el 35% de las cuentas del programa. Esta regla confirma la validez de los clusters.



Segmento	% de Puntos Totales	% de Cuentas
Top	62%	19%
Loyal	18%	16%
Medium	13%	27%
Light	4%	22%
New	0%	2%
Descarte	3%	14%
	100%	**100%**

El 35% de las Cuentas acumula el 80% de los Puntos Totales



A su vez, la descripción de las variables originales para cada uno de los segmentos muestra claramente las diferencias de comportamiento.

	Variables Descriptivas por Cuartiles			
	Segmentos	**P25%**	**P50%**	**P75%**
Índice Multimarca	Top	2	4	5
	Loyal	2	3	5
	Medium	1	2	3
	Light	1	2	2
	New	1	1	2
Recency	Top	12	18	44
	Loyal	14	19	30
	Medium	22	68	351
	Light	26	143	462
	New	15	22	31
Puntos standard por mes	Top	126	166	247
	Loyal	55	67	82
	Medium	27	35	45
	Light	7	11	16
	New	26	50	95
Frecuencia (dias entre transacciones)	Top	5	7	13
	Loyal	7	10	13
	Medium	16	22	32
	Light	28	43	73
	New	6	10	22

Definición de los Segmentos a partir de las variables descriptivas

	Variables Descriptivas por Segmento			
P50%	**Índice Multimarca**	**Recency**	**Puntos Standard por mes**	**Frecuencia**
Top	**4**	**18**	**166**	**7**
Loyal	**3**	**19**	**67**	**10**
Medium	**2**	**68**	**35**	**22**
Light	**2**	**143**	**11**	**43**
New	**1**	**22**	**50**	**10**

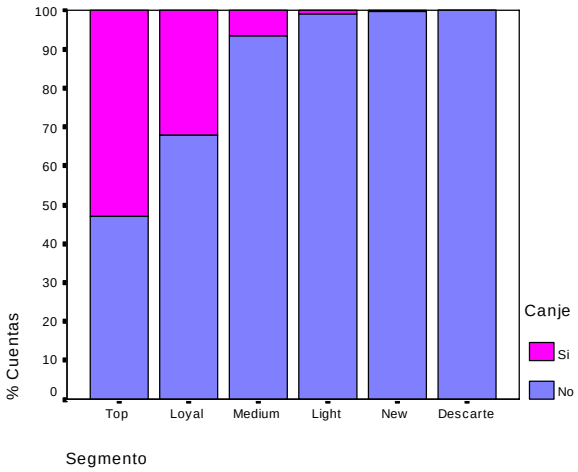
A partir de la segmentación, los distintos segmentos se describen a partir de sus categorizaciones en las variables descriptivas, por ejemplo:

Un cliente Top (P50%) es un cliente que:

- ☑ **Compra en 4 Empresas Participantes del Programa.**
- ☑ **La última transacción la realizó hace 18 días[4].**
- ☑ **Acumula 166 puntos Standard por mes.**
- ☑ **Cada 7 días realiza una compra en las Empresas Participante.**

Como puede observarse, los segmentos Top y Loyal presentan similares patrones de uso de la tarjeta, medido en su composición de Frecuencia, Recencia, e índice multimarca, pero con una marcada diferencia en la tasa de acumulación de puntos por mes.

Análisis del Canje de Recompensas -



La aplicación de la segmentación permitió determinar que más del 50% de los clientes Top había canjeado al menos una recompensa a Diciembre de 2001. La evolución de este indicador se analizará a lo largo del tiempo.

Las cuentas canjeadoras se concentraban en los segmentos Top y Loyal.- Esta conclusión confirma el modelo de segmentación ya que los mejores clientes son los que más valoran el programa.

[4] El Recency se definió como la diferencia en días entre la Fecha del estudio y la Fecha de la última transacción de la Cuenta. El Recency es una variable que depende de la "fecha de corte" para el análisis y no así la Frecuencia.

Actuar a partir de los resultados

Aplicación del Modelo de Consumos Travelpass

EL CRM es el proceso que administra la relación entre las empresas y sus clientes.
Para que resulte exitoso es necesario segmentar la base de clientes en función de sus patrones y perfil de comportamiento.

Sobre esta base conceptual, se planteó como uno de los principales objetivos del área de Marketing, el diseño de la **Estrategia de Relacionamiento con los Clientes,** a partir de la segmentación de Clientes Travelpass.

Objetivos del Ciclo de Relacionamiento con los Clientes

- **Construir un vínculo sólido, rentable y duradero entre los clientes y Travelpass.**
- **Sistematizar el programa de relacionamiento.**
- **Facilitar la medición de performance de las comunicaciones.**

Construir un vínculo sólido…

- ✓ **Para afianzar el valor de la marca, como fuente de satisfacciones, en estrecha relación con las Empresas Participantes.**
- ✓ **Para construir barreras de salida del Programa mediante el aporte de valor agregado.**
- ✓ **Para estimular el consumo de los productos y servicios de las Empresas Participantes y de esa manera mejorar su negocio.**

El Modelo de Segmentación como Soporte de la Estrategia de Relacionamiento

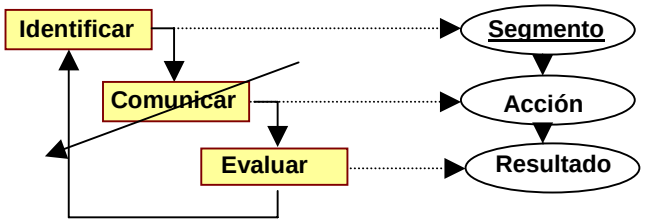
El Modelo de Segmentación se aplicó en todas las áreas de Relacionamiento con el cliente permitiendo re-asignar los recursos humanos y tecnológicos conforme al nivel de atención y comunicación requerido para cada segmento.

- MARKETING Y COMUNICACIONES
- CENTRO DE ATENCIÓN TELEFÓNICA
- INTERNET
  - DESARROLLO DEL CANAL
  - EXPLORACIÓN DE LOS SEGMENTOS (ENCUESTAS E INVESTIGACIONES DE MERCADO)

- MARKETING Y COMUNICACIONES

Las Comunicaciones del Programa constituyen el principal vínculo hacia el Cliente Travelpass.

La Estrategia de Comunicaciones se planteó en función del perfil de cada segmento teniendo en cuenta:



- **Identificar el segmento objetivo, sobre la base del Modelo de Segmentación.**

- **Definir el objetivo de comunicación** según la etapa de relacionamiento y generar la acción de Database Marketing:CaptaciónInducciónReconocimientoUp-GradingRetenciónRecuperación
- **Evaluar y medir el resultado de las acciones**

Matriz de Relacionamiento con los Clientes - Comunicaciones proactivas

Segmento	Etapa de Relacionamiento con los Clientes					
	Captación	Inducción	Reconocimiento	Up-Grading	Retención	Recuperación
TOP	Acción Member Get Member		▪ Resumen de Puntos ▪ Saludo de cumpleaños ▪ Encuesta de satisfacción ▪ Beneficios especiales	▪ Incentivo del cross traffic ▪ Saludos en eventos especiales ▪ Promociones especiales	▪ Canje de recompensas ▪ Cross Traffic ▪ Acumulación de Puntos ▪ Promociones	
LOYAL			▪ Resumen de Puntos	▪ Incentivo del cross traffic ▪ Saludos en eventos especiales ▪ Promociones especiales	▪ Cross Traffic ▪ Acumulación de Puntos ▪ Promociones	▪ Promociones
MEDIUM				▪ Acciones para incentivar Acumulación		
LIGHT				▪ Acciones para incentivar Acumulación		
NEW		▪ Resumen de Puntos ▪ Acción de Bienvenida ▪ Encuesta de Satisfacción				

- ☑ **Todos los esfuerzos de relacionamiento (comunicación, atención personalizada y beneficios) se centralizaron en los segmentos Top y Loyal del Programa**

- ☑ **Estos dos segmentos representaban el 35% de las Cuentas que acumulaban el 80% de los Puntos Totales.**

- ☑ **Ambos segmentos constituyeron el target de las acciones desarrolladas en el Programa y en las Empresas Participantes.**

ESTRATEGIA DE RELACIONAMIENTO CON LOS CLIENTES TOP

Perfil de los Clientes

Clientes que manifiestan un comportamiento muy comprometido con el Programa y las Empresas Participantes: son los clientes más fieles al Programa.

- **Poseen una alta acumulación de Puntos**: 166 Puntos Travelpass por mes, equivalente -en promedio- a $ 330.- de compras en las Empresas Participantes.
- **Manifiestan una Alta Frecuencia**: cada 7 días presentaban su Tarjeta Travelpass en alguna de las Empresas Participantes del Programa.
- **Baja Recencia** lo cual marca un elevado nivel de actividad en los últimos días.
- Y el **Índice Multimarca más alto** entre la base de clientes: lo cual significa que adhirieron a un elevado número Empresas Participantes del Programa.

Estrategia de Relacionamiento

Objetivo: **Reconocer su Fidelidad en el Programa a través de acciones específicamente diseñadas para ellos.**

Fundamento Estratégico:

El reconocimiento a los clientes es fuente de satisfacción y anclaje. Permite establecer y afianzar vínculos más personalizados.

- **Acción de Cumpleaños de Clientes con Beneficios especiales para disfrutar en el mes.**
  - Los beneficios consisten en Cenas en La Caballeriza, regalos o descuentos de las Empresas Participantes u otros regalos (por ejemplo, botellas de champagne con un par de copas).
  - Mediante Encuestas de Satisfacción telefónicas se monitorea el impacto de la acción en los clientes, arrojando **índices de satisfacción cercanos al 100%.**
  - Mediciones posteriores manifestaron que los clientes que reciben el reconocimiento de cumpleaños incrementan en un **10% su Tasa de Acumulación de Puntos**.
  - También se reciben llamados e e-mails de agradecimiento por la acción.
  - Colateralmente existe **un beneficio concreto para las Empresas Participantes** del Programa, que ven incrementados sus índices de captación ó retención de clientes, con clientes que poseen un ticket promedio mayor a la media de los clientes.
- **Reconocimientos especiales para las fechas especiales: aniversario de su activación en el Programa (12 meses de permanencia en el Programa); aniversario del lanzamiento del Programa**
  - Los beneficios consisten en Cenas, Avant Premiers y Entradas de Cine obsequiadas por el Programa.
  - Mediciones posteriores manifestaron que éstos clientes permanecen en el Programa más tiempo que los clientes que no reciben el reconocimiento (para ello se realizan pruebas de testeo con grupos de control monitoreados en el tiempo.
- **Comunicación permanente a través del Resumen de Puntos Travelpass.**
  - Los clientes Top constituyen el target principal de toda la comunicación generada por el Programa: catálogos de recompensas, promociones segmentadas y beneficios especiales de las Empresas Participantes, resúmenes de Puntos.
    - Durante el año 2000, la comunicación de Travelpass se realizó con una frecuencia bimestral.
    - Durante el año 2001, la frecuencia de envíos se redujo a trimestral.

Objetivo: Desarrollar al máximo la capacidad de consumo de éstos clientes en el Programa, incrementando su índice multimarca (cross-traffic) y los Puntos acumulados por mes.

- **Incentivar el Cross-selling y Cross – Traffic**

  - Para estos clientes se diseñaron acciones especiales de Puntos Extra y Beneficios exclusivos para incentivar el cross traffic en las Empresas Participantes.

    Por la atractividad del perfil, se hace mucho hincapié en que los Clientes Top compren en los cuatro socios del Programa, esto significa que posean un índice multimarca igual a 4 en los socios. Para ello se desarrollan acciones especiales con importantes incentivos: Por ejemplo: si el cliente compra en Norte pero no compra en Shell, se les envía una pieza de marketing directo con una promoción exclusiva al efecto.

    Los eventos especiales también sirven como disparador para el incremento del cross selling. Acciones de Marketing Directo realizadas con motivo del Día del Niño, Día del Padre o Día de la Madre funcionan como disparador para el incremento del cross-selling en los Clientes Top.
    En estas acciones se obtienen **Tasas de Respuesta del orden del 10%.**

Objetivo: Retener a los Clientes más fieles del Programa, a través del incentivo del canje de recompensas.

Fundamento Estratégico.

Los clientes que canjean recompensas "experimentan y tangibilizan" la promesa del Programa. Ello genera mayor satisfacción y anclaje.
Por eso es "deseable" que canjeen los mejores clientes del Programa.

- Para estos clientes se diseñan acciones tendientes a Incentivar el Canje de Recompensas. Se envían promociones especiales y descuentos de puntos en recompensas exclusivas.
- El estímulo personalizado de canje ha logrado incrementar en 5 puntos el nivel de canje de éstos clientes en un mes determinado.
- Se diseñó un Catálogo con recompensas "aspiracionales" para incentivar la acumulación de puntos en el tiempo.
- También se envían promociones especiales y descuentos de Puntos en recompensas exclusivas.

ESTRATEGIA DE RELACIONAMIENTO CON LOS CLIENTES LOYAL

Perfil de Clientes

Clientes que manifestaban un comportamiento muy comprometido con el Programa y las Empresas Participantes, pero con una Tasa de Acumulación de Puntos menor.

- **Tasa de Acumulación de Puntos**: 67 Puntos Travelpass por mes, equivalente al 40% de los puntos sumados por un Cliente Top.
- **El resto de las variables de comportamiento, mantienen un estado similar al de un Cliente Top.**

Estrategia de Relacionamiento

Objetivo: Reconocer su fidelidad en el Programa a través de una comunicación permanente y estable.

- **Comunicación permanente a través del Resumen de Puntos Travelpass.**
  - Los clientes Loyal constituyen el target principal de toda la comunicación generada por el Programa: catálogos de recompensas, promociones segmentadas y beneficios especiales de las Empresas Participantes, resúmenes de Puntos.
    - Durante el año 2000, la comunicación de Travelpass se realizó con una frecuencia bimestral.
    - Durante el año 2001, la frecuencia de envíos se redujo a trimestral.

Objetivo: Desarrollar al máximo la capacidad de consumo de éstos clientes en el Programa, incrementando su índice multimarca (cross-traffic) y los Puntos acumulados por mes.

- **Incentivar el Cross-selling y Cross – Traffic**
  - Los clientes Loyal poseen un alto potencial de pasar al segmento de Clientes Top si se los incentiva con acciones de Cross selling que les permitan incrementar su ticket y/o consumo promedio mensual.

    Para estos clientes se diseñaron acciones especiales de Puntos Extra y Beneficios exclusivos para incentivar el Cross selling y la cantidad de empresas participantes en las que compraban.

    El 3% de los clientes Loyal pasaron a ser Clientes Top durante el año 2001.

Objetivo: Recuperar clientes con una recencia mayor a 3 meses.

- Los clientes Loyal con una Recencia mayor a 3 meses en el Programa son incentivados con promociones especiales y beneficios o descuentos.
- Al siguiente mes, se vuelve a formular el estímulo, sólo si se posee e-mail en la base Travelpass.

ESTRATEGIA DE RELACIONAMIENTO CON LOS CLIENTES MEDIUM

Perfil de Clientes

Clientes que manifiestan muy bajo potencial y compromiso con el Programa.

- **Tasa de Acumulación de Puntos**: 35 Puntos Travelpass por mes, equivalente al 50% de los puntos sumados por un Cliente Loyal.
- **Frecuencia:** 22 días
- **Recencia:** 68 días.
- **Índice Multimarca**: 2. La media de los clientes Medium compra en dos Empresas Participantes del Programa.

Estrategia de Relacionamiento

Objetivo: Incentivar el Cross-selling, Cross – Traffic y la Tasa de Acumulación de Puntos de los clientes en el Programa

- Para estos clientes se diseñaron acciones especiales de Puntos Extra y Beneficios exclusivos para incentivar el Cross selling, la cantidad de empresas participantes en las que compraban y la Tasa de Acumulación de Puntos Travelpass.
- En líneas generales la respuesta de éstos clientes a los incentivos de comunicación no es lo suficientemente atractiva como la insistir en aplicaciones posteriores.

ESTRATEGIA DE RELACIONAMIENTO CON LOS CLIENTES LIGHT

- El Segmento de Clientes Light fueron los menos atractivos para el Programa.
- En estos clientes no se invirtió esfuerzo en comunicaciones periódicas o programadas ya que no constituían el "core" de clientes interés para el Programa. Durante los primeros tiempos se los indujo con acciones puntuales para incrementar la tasa de acumulación de puntos, pero los intentos fueron incipientes y estos clientes poco a poco fueron dados de baja.

ESTRATEGIA DE RELACIONAMIENTO CON LOS CLIENTES NEW

Objetivo: Inducir a los clientes nuevos en el Programa y estimular su acumulación de Puntos a través del incentivo de cross-selling y cross traffic

- **Comunicación permanente a través del Resumen de Puntos Travelpass.**
  - Para estos clientes se desarrolló una Acción de "Bienvenida al Programa" con la descripción del Programa, las Recompensas y las Empresas Participantes.
  - Durante los primeros cuatro meses de Activación los clientes New constituyen el target principal de toda la comunicación generada por el Programa: catálogos de recompensas, promociones segmentadas y beneficios especiales de las Empresas Participantes, resúmenes de Puntos.

- MARKETING Y COMUNICACIONES – ENVÍOS DE RESÚMENES DE PUNTOS Y CATÁLOGOS

Durante los primeros años del Programa se envió Resúmenes de Puntos e información promocional con una frecuencia bimestral a toda la base de clientes Travelpass.

En esa primera etapa, las características del envío eran las siguientes:

- Envío con Resumen de Cuenta y Catálogo de Recompensas masivo.
- Promociones masivas (no segmentadas).
- Revista de contenido general.

A partir Julio de 2000, se aplicaron los principios de la segmentación, permitiendo maximizar la tasa de respuesta de los envíos y promociones y optimizar los costos operativos asociados.

Aplicando la segmentación de clientes fue posible dirigir mensajes segmentados (DBM), promociones y beneficios según el perfil del cliente, a la vez que se redujeron la cantidad de envíos a los segmentos de mayor valor para el programa.

El target principal de los envíos de Resúmenes de Puntos y Catálogos está constituido por los clientes Top y Loyal Travelpass.

Resultados obtenidos:

La aplicación de la segmentación de clientes permitió obtener tasas de respuesta del orden del 10% en acciones de Marketing Directo.

A partir de la implementación de los envíos segmentados se produjeron importantes ahorros asociados al 64% de las cuentas (Medium, Light y Descarte) a las que se les dejó de enviar el resumen bimestral, las cuales representaban el 19% de los Puntos Totales acumulados.

- CENTRO DE ATENCIÓN TELEFÓNICA

El Centro de Atención Telefónica es un servicio tercerizado desde el principio del Programa.

En el IVR se habilitó la atención automatizada de consultas a través de la cual, previa identificación del Cliente, se puede acceder a un árbol de opciones de preguntas-respuestas frecuentes: recompensas disponibles, vencimiento de puntos, saldo de puntos, etc.

Para el caso de la **atención personalizada** (40% del total de llamadas ingresadas al call center), una vez filtrado el IVR, se aplicó modelo resultante de la segmentación de clientes a efectos de garantizar la mejor atención en los clientes pertenecientes a los segmentos Top y Loyal y la optimización de los costos operativos de atención.

Se introdujeron "**colas de atención diferenciadas**" de acuerdo al perfil del cliente:

- Prioridad Alta: Segmentos Top / New.
- Prioridad Media: Segmentos Loyal
- Prioridad Baja: restantes segmentos

Segmento	Prioridad	Nivel de Atención	Tiempo de espera promedio (para la atención personalizada)
Top / New	**Alta**	**97%**	**23"**
Loyal	**Media**	**91%**	**35"**
Resto de los segmentos	**Baja**	**75%**	**45"**

Base de actividad: 140.000 llamadas IVR / 42.000 requirieron atención personalizada (julio 2001)

Resultados obtenidos:

La aplicación del Modelo de Segmentación garantiza un óptimo nivel de atención para los clientes más valiosos del Programa, en un adecuado equilibrio y relación de costos.

- INTERNET – DESARROLLO DEL CANAL Y EXPLORACIÓN DE LOS SEGMENTOS

El site de Travelpass ha ido evolucionando desde sus comienzos, migrando desde una perspectiva meramente informativa a un enfoque transaccional e interactivo, en marzo de 2001.

La ampliación de las posibilidades de interacción con el cliente abrió un nuevo canal de comunicación. Se desarrollaron acciones y promociones incentivando la registración de los clientes y el canje de recompensas a través de la web.

Caso 1: REGISTRACIÓN DE CLIENTES

En Julio de 2001 se desarrolló una acción para incentivar la registración de clientes en el site y explorar el perfil sociodemográfico y psicográfico de los clientes más valiosos del Programa:

- **Acción:** Reconocimiento a clientes con que compraran en los 4 socios del Programa (Índice Multimarca = 4 –en socios-).

- **Objetivos:**

✓ Incentivar la registración en el web site.
✓ Optimizar la calidad de información de los mejores clientes Travelpass
✓ Establecer una nueva vía de comunicación por Internet que redundará en beneficios económicos y prácticos.

- **Target:**

✓ Clientes con Índice Multimarca igual a 4 en los socios, esto es que compraran en Shell y Norte y tuvieran servicios adheridos a Telecom y a Banco Galicia.

- **Universo contactado:** 9.200 Clientes.

- **Tasa de respuesta**: **23%** (medida en términos de incremento de clientes registrados)

- **Descripción de la acción:**

  Para incentivar la registración en el site se ofreció a los clientes “Multimarca 4 en los Socios” una recompensa por ingresar al Web Site: www.travelpass.com.ar, registrarse y responder algunas preguntas referidas a las empresas participantes. El incentivo eran Puntos Extra.

  Luego se analizó el perfil sociodemográfico y de segmento de los respondentes:

  - **El 90% pertenecía al segmento Top Travelpass.**
  - **El 46% indica como ocupación “Profesional”.**
  - **El 59% son varones.**
  - **La edad promedio del respondente es de 40 años.**
  - **El grupo familiar promedio tiene 3,4 integrantes.**

Una vez más se confirma que la segmentación es efectiva para identificar a los mejores clientes del programa.

Caso 2: ENCUESTA GRUPO TELECOM: ¿Cómo obtener la mejor respuesta?

El éxito obtenido en la primer experiencia de investigación en el site impulsó al Grupo Telecom a promover entre sus clientes una encuesta sobre productos y servicios de telefonía residencial, celular y servicios de Internet. **La acción se desarrolló en Diciembre de 2001.**

- **Acción:** Encuesta Grupo Telecom en el site de Travelpass
- **Objetivos:**
  - **Grupo Telecom**

    ✓ Recopilar datos relacionados con los servicios de telefonía y perfil sociodemográfico de los clientes, a fin de diseñar propuestas de servicios ad hoc.
    ✓ Vincular el perfil de consumo con el perfil sociodemográfico de los clientes, a efectos de potenciar el verdadero conocimiento sobre los mismos.

  - **Travelpass**

    ✓ Incentivar la registración de clientes en el site, mediante acciones de valor agregado.
    ✓ Evaluar el impacto y la tasa de respuesta de la comunicación física vs. la comunicación por Internet.
- **Target:**

✓ Clientes pertenecientes a los Segmentos Top y Loyal en el Programa Travelpass y Segmento "4-Muy Buenos" en Telecom (según Modelo de Segmentación de Clientes Telecom).

- **Universo contactado:** 78.104 clientes.
  - **Canales de comunicación**:
    La comunicación se realizó mediante un envío físico y un envío adicional por e-mail a los clientes registrados en la base Travelpass

    - Todos los clientes invitados a responder la encuesta (78.104) recibieron el envío postal.
    - Los clientes del Target registrados en el Web Site 23% del total (17.968) recibieron previamente un e-mail.

- **Incentivo**: Puntos Extra
- **Tasa de respuesta general**: **11 %.** (**50%** fueron clientes registrados con motivo de la acción)

- **Descripción de la acción:**

  Para incentivar la participación en la Encuesta del Grupo Telecom en el site de Travelpass se ofreció a los clientes una recompensa por ingresar al Web Site: www.travelpass.com.ar, registrarse y completar la encuesta.

  La encuesta constaba de cuatro módulos: **Telefonía residencial**, Telefonía **Celular**, **Internet**, y **Perfil Grupo Familiar**.

- **Características de la Encuesta**:

  - Todas las preguntas eran de **respuesta obligatoria**.
  - El avance a la pregunta ó módulo siguiente dependía de la consistencia de la pregunta / módulo anterior.
  - Todas las preguntas **eran cerradas**.
  - El perfil de las preguntas era de satisfacción, modalidad de uso de los productos o servicios, hábitos y preferencias.
  - Tiempo estimado para la respuesta **5 a 7 minutos**.
  - Cantidad de preguntas a responder: **35 preguntas**

- **Pasos para acceder a la encuesta:**

  1. **Ingresar a www.travelpass.com.ar**
  2. **Registrarse en el site**
  3. **Completar la Encuesta del Grupo Telecom**

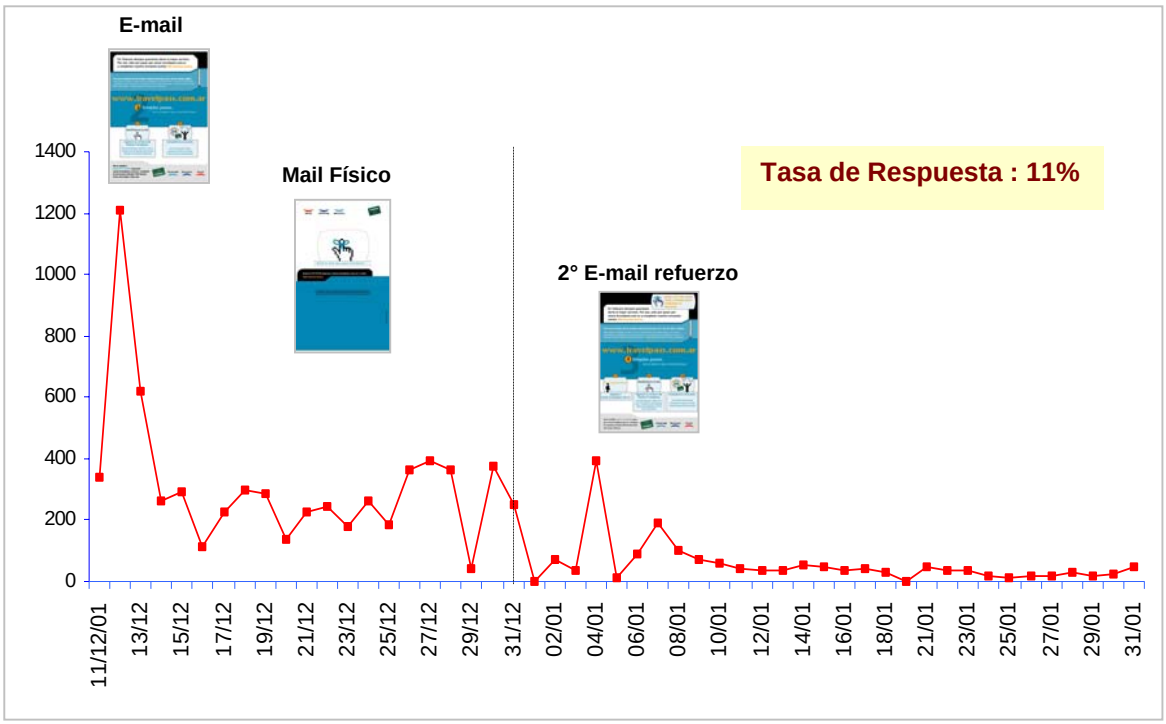
  Luego se analizó la tasa de respuesta y el perfil sociodemográfico y de segmento de los respondentes.

- **Análisis de la Tasa de Respuesta:**

  Resultó interesante analizar la tasa de respuesta en relación con los dos canales de comunicación.

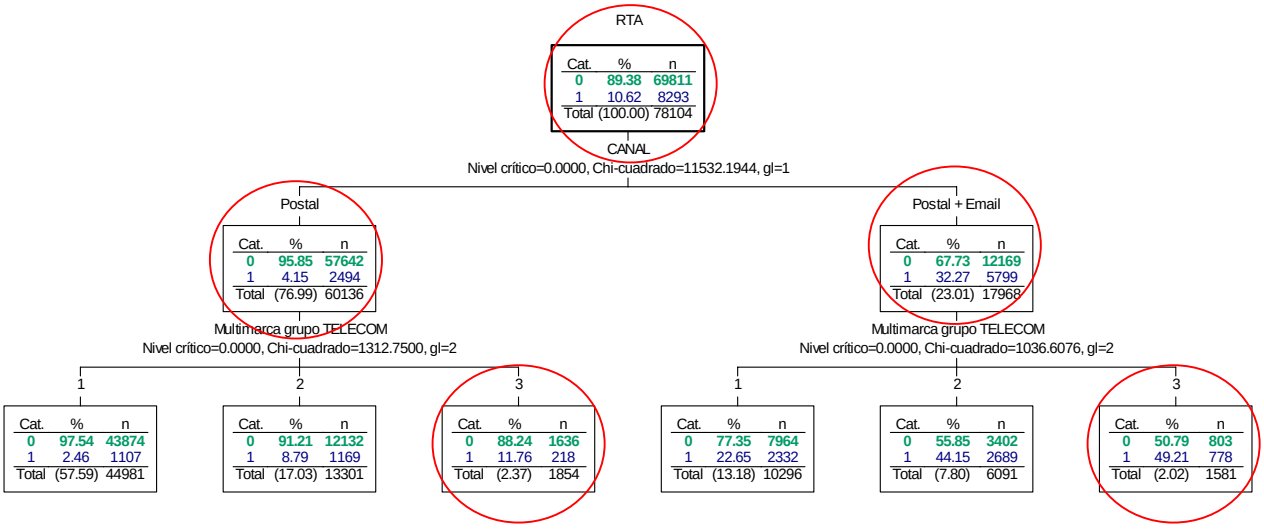
  El gráfico muestra la evolución de la respuesta en el tiempo, donde puede verse que el **30% de las respuestas** se obtuvo en los 3 días posteriores al envío del email. Ello pone de manifiesto el alto e inmediato impacto de respuesta de las acciones por este canal.

Evolución de la Tasa de Respuesta de la Encuesta



Resulta interesante además, el análisis obtenido mediante un árbol de segmentación CHAID del perfil y tasa de respuesta.

Árbol de Decisión CHAID – ANÁLISIS DE LA TASA DE RESPUESTA DE LA ENCUESTA



Para explicar la tasa de respuesta se utilizaron variables de consumo obtenidas del datawarehouse y no del cuestionario.

- La tasa de respuesta de los que recibieron solamente la pieza postal (POSTAL) fue del **4,15%**. En el grupo que recibió el email (POSTAL-Email) en cambio fue del **32,27 %.**
- El siguiente predictor es la cantidad de empresas del Grupo Telecom en las que compra el cliente.
- Para los clientes de las 3 empresas del Grupo – Telecom (telefonía residencial), Personal (telefonía celular) y Arnet (servicios de Internet) – la tasa de respuesta sube al **49,21%.**
- Como siguientes predictores aparecen la profesión y el segmento Travelpass.

Estas primeras conclusiones han promovido ya como primer objetivo del año 2002 la promoción de una campaña de registración masiva de clientes en el site.

Desde Diciembre de 2001 **Internet se ha convertido en el principal y único canal de comunicación para los clientes Travelpass**. Los envíos físicos –resultantes de las acciones comprendidas en el Ciclo de Relacionamiento- se han reducido a un contacto por año por cliente e Internet pasó a ser el canal de comunicación por excelencia.

Esta decisión requiere un gran esfuerzo por parte de Travelpass, para incentivar a sus mejores clientes a registrarse en el site.

La registración de los clientes

Los clientes registrados reciben una Clave de Identificación Personal que los acredita para la realización de transacciones (canjes) en la web.-

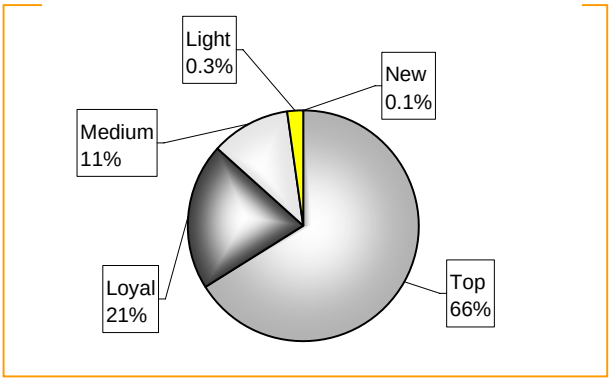
El Plan de comunicaciones vía Internet incluye:

- **Acciones de captación de e-mails**
- **Acciones de Incentivo a la registración de clientes, dirigidas principalmente a los clientes Top y Loyal**
- **Incentivo del canje de recompensas a través de la web.-**
- **Envíos de comunicaciones y promociones segmentadas.**

Los perfiles de clientes también se aplican para segmentar el universo de envíos de e-mails por Internet.

Tal como puede observarse en el gráfico, el envío de los e-mails funciona como “call to action” para la afluencia de clientes a la web y el incremento de clientes Registrados.-

Segmentación de Clientes que canjean recompensas en Internet



La segmentación de clientes también se aplica para analizar el perfil de los clientes activos en Internet.

El 50% de los canjes del Programa se realizan a través de Internet y como puede observarse, los clientes más valiosos (Top y Loyal) son los que más adhieren a esta modalidad de canje.

Una vez más se confirma que el modelo de segmentación es efectivo para identificar a los clientes más valiosos del programa.

MODELOS DE SOCIOS TRAVELPASS

Promediando el año 2001 se inició un trabajo de segmentación sobre los clientes de la base Travelpass que compran productos o adhirieron servicios en los socios del Programa: Supermercados Norte, Shell, Grupo Telecom y Banco Galicia.

El propósito del estudio fue desarrollar Modelos de Segmentación de los clientes Travelpass según su comportamiento de consumo en la empresa socia, para identificar a los mejores clientes, sus características en Travelpass y orientar la búsqueda de mellizos.

Modelo de Supermercados Norte: El Ticket Óptimo en el Supermercado

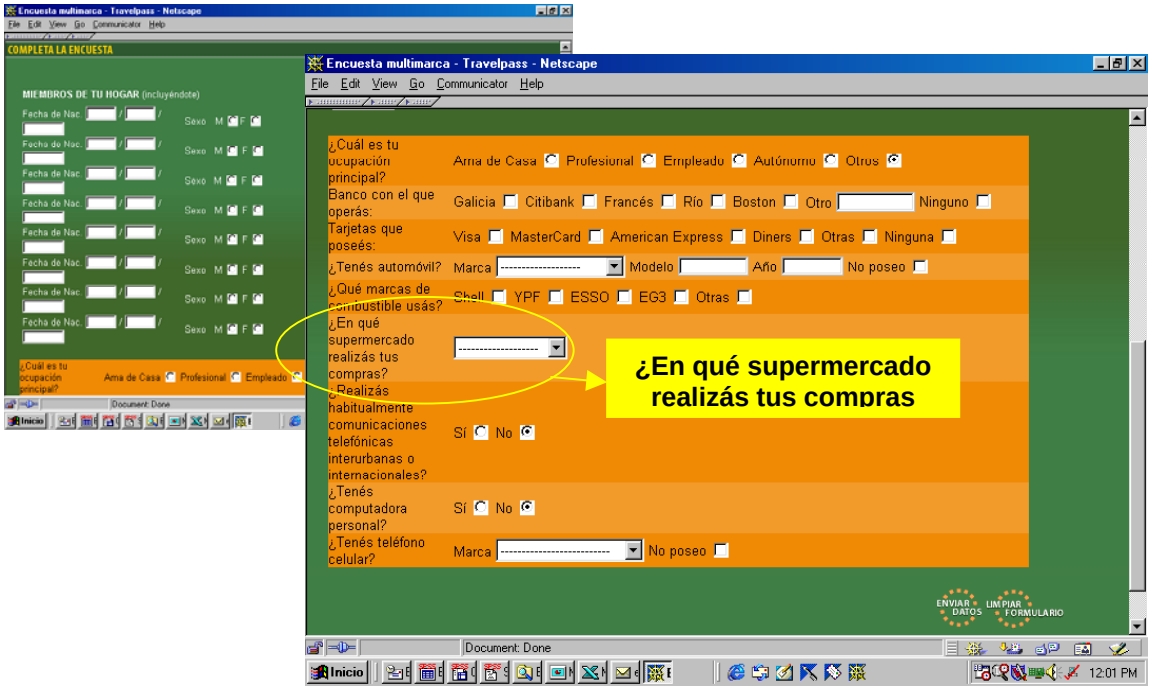
La acción de incentivo de registración que se realizó sobre los clientes " Multimarca 4 en socios" sirvió como disparador para desarrollar el Modelo de Socios de Supermercados Norte.

Como se describió anteriormente, para incentivar la registración en el site se ofreció a los clientes una recompensa por ingresar al Web Site: www.travelpass.com.ar, registrarse y **responder algunas preguntas referidas a las empresas participantes**.

Luego se analizó el perfil sociodemográfico y de segmento de los respondentes:

- **El 90% pertenecía al segmento Top Travelpass.**
- **El 46% indica como ocupación "Profesional".**
- **El 59% son varones.**
- **La edad promedio del respondente es de 40 años.**
- **El grupo familiar promedio tiene 3,4 integrantes**

DISEÑO DE LA ENCUESTA EN EL SITE



EL TICKET OPTIMO EN EL SUPERMERCADO

Entre otras preguntas de la encuesta figuraba como pregunta de opción simple:

¿En qué supermercado realizas tus compras?

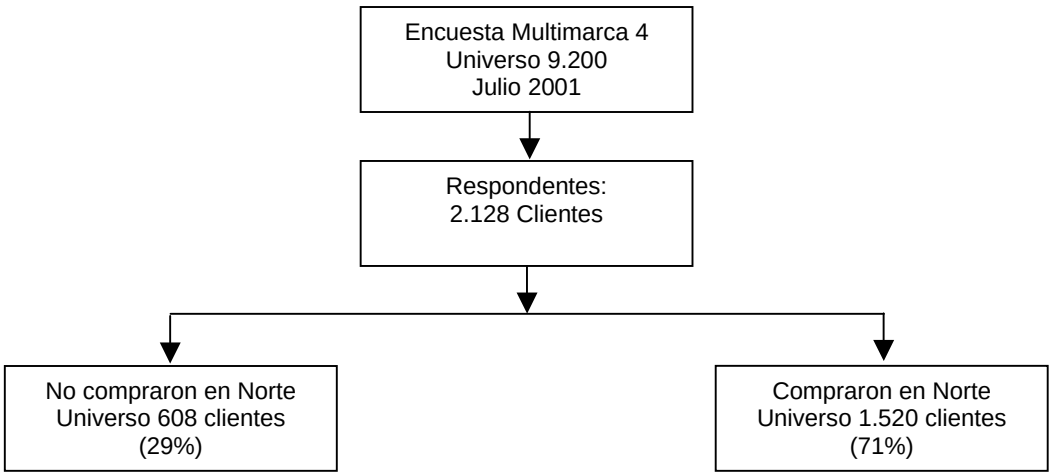
A pesar de que todos los entrevistados tenían transacciones en Norte durante el año 2001, solamente el **67% lo identificó como su supermercado habitual**.

Supermercado	Cuentas	Porcentaje
Norte	1424	67%
Coto	210	10%
Carrefour	198	9%
Jumbo	133	6%
Disco	125	6%
Wal Mart	29	1%
Ekono	5	0%
Auchan	4	0%
Total	2128	100%

Para analizar el comportamiento en el supermercado se tomó del Datawarehouse Travelpass las transacciones en NORTE de las cuentas que respondieron la encuesta (finalizada el 31 de julio) en los meses de Agosto y Septiembre.

Del total de clientes que respondieron:

- **1520 clientes (71%) compró en esos meses.**
- **608 clientes (29%) NO compró en esos meses**
- **La compra mensual promedio fue de $ 195.**



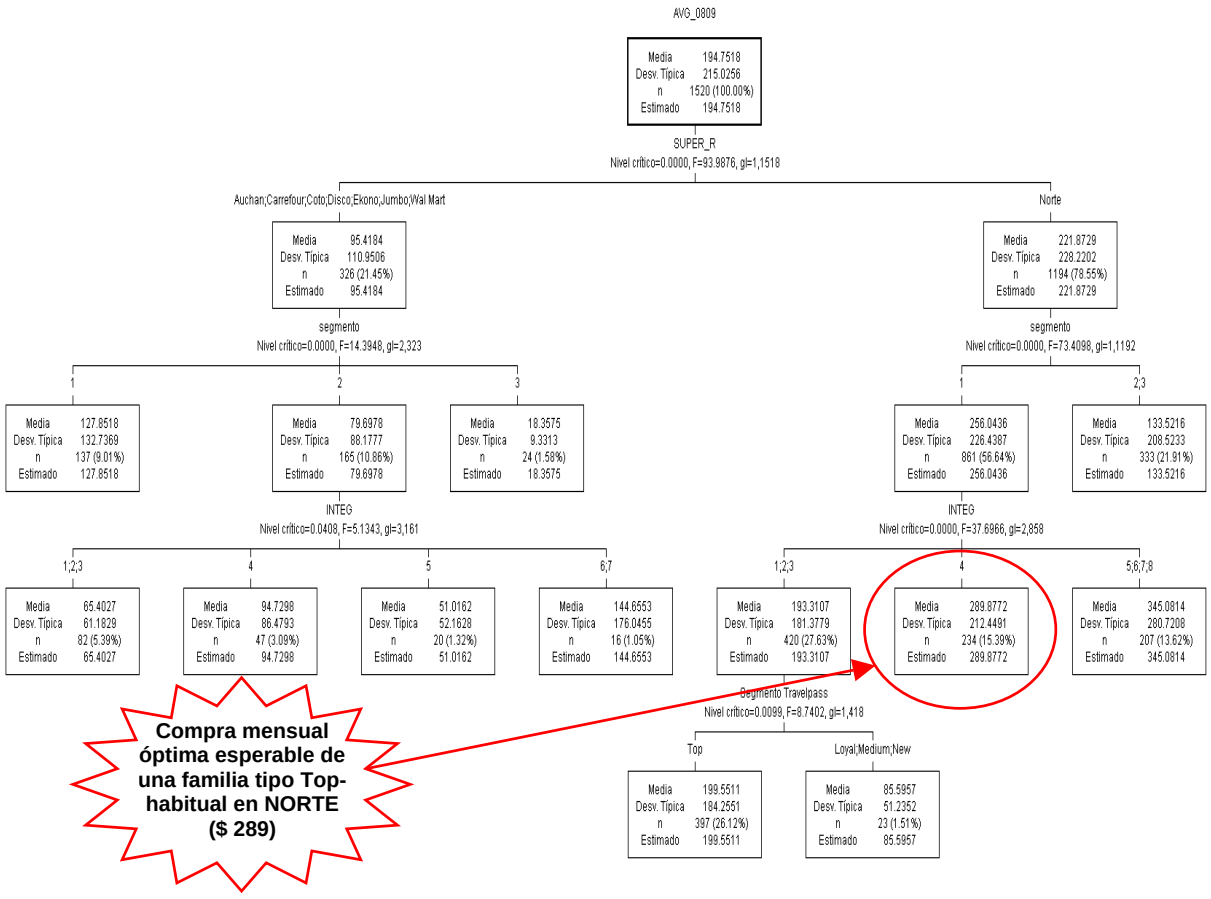
Para los que no compraron en los meses de Agosto y Septiembre, la composición de marcas era la siguiente:

Supermercado	Cuentas	Porcentaje
Norte	230	38%
Coto	125	21%
Carrefour	99	16%
Jumbo	71	12%
Disco	64	11%
Wal Mart	16	3%
Ekono	2	0%
Auchan	1	0%
TOTAL	608	100%

La **compra mensual promedio de $ 195. -** varía si se la analiza separando a las cuentas según diferentes variables. Este estudio de segmentación se realizó con un árbol de decisión CHAID que muestra los grupos para los cuales existen diferencias estadísticamente significativas para la variable objetivo (consumo mensual).

Al monto mensual promedio se agregaron otras variables como: integrantes del grupo familiar, segmento en NORTE, segmento en Travelpass, etc.

Árbol de Decisión CHAID - LA COMPRA MENSUAL EN EL SUPERMERCADO



Es importante destacar que:

- La compra mensual promedio de $ 195 sube a $ 222 para los clientes que indican a NORTE como "el supermercado en donde realiza sus compras" y baja a $ 95 para los restantes.

- El segmento propio de NORTE, y la cantidad de integrantes del grupo familiar son los otros predictores del consumo mensual.

El valor encontrado entonces en el nodo determinado por los clientes habituales de Norte que tienen un grupo familiar de 4 integrantes se tomó entonces como "consumo óptimo mensual" y corresponde a $ 289 por mes.

Con este valor hallado por el árbol para el "consumo óptimo mensual" más la consideración de la **recencia en NORTE** se procedió a segmentar a los clientes del supermercado Norte.

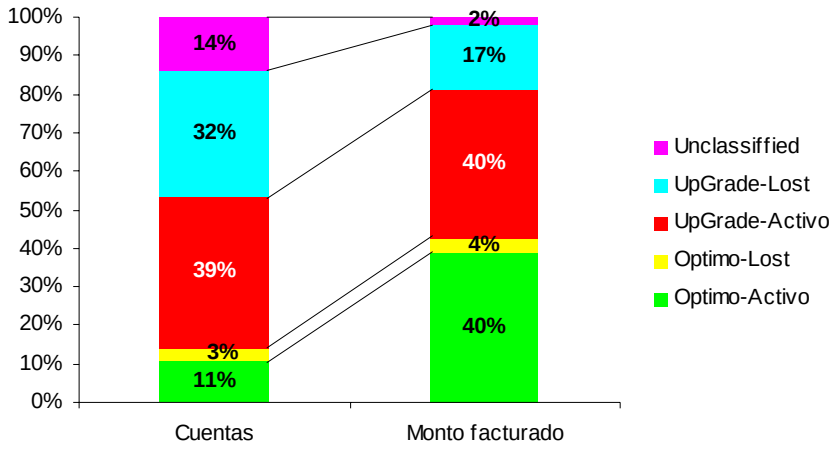
MODELO DE SOCIOS "SUPERMERCADOS NORTE" EN TRAVELPASS

A partir de este estudio se definieron 4 segmentos con las siguientes características

Segmento	Característica
Optimo – Activo	Consumo mensual igual o superior al óptimo esperado, fecha última compra hasta 60 días
Optimo – Lost	Consumo mensual igual o superior al óptimo esperado, fecha última compra mayor a 60 días
Upgrade – Activo	Consumo mensual inferior al óptimo esperado, fecha última compra hasta 60 días
Upgrade – Lost	Consumo mensual inferior al óptimo esperado, fecha última compra superior a 60 días

Esta segmentación – muy sencilla de calcular – resulta sumamente efectiva ya que conduce a un Diagrama de Pareto de participación en la facturación en donde se observa que el **11% de las cuentas (Optimo – Activo) acumula el 40% de la facturación**.

MODELO DE SEGMENTACIÓN SUPERMERCADOS NORTE
Relación Segmento - % de Cuentas – Monto Facturado



MODELOS DE SOCIOS TRAVELPASS: Grupo Telecom

Para el Modelo de Socios del Grupo Telecom se aplicó una metodología de trabajo similar a la descripta en el proceso de Segmentación General de Travelpass.

Modelo de Segmentación para el Grupo Telecom

1. Marco conceptual

- Se trataba de un problema de clustering, ya que se desconocía a priori la conformación de los segmentos y cuáles son las variables que discriminaban el comportamiento.
- La metodología a emplearse tuvo un enfoque "data driven", es decir sin supuestos ni restricciones "a priori".

2. Disponibilidad de Datos

- Se disponía de las transacciones de los clientes Travelpass que consumían servicios del Grupo Telecom.
- A partir de los datos originales se calculan variables derivadas, como por ejemplo el promedio mensual de puntos Standard y la recencia.

3. Variables de análisis

Para cada cuenta se dispuso de las siguientes variables:

- **Puntos Std_mes en Telecom**
- **Puntos Std_mes en Telecom Personal**
- **Puntos Std_mes en Telecom Internet**
- **Recencia**: días desde la última acreditación de puntos en el Grupo Telecom.
- **Activación**: días desde la primera transacción en el Grupo Telecom

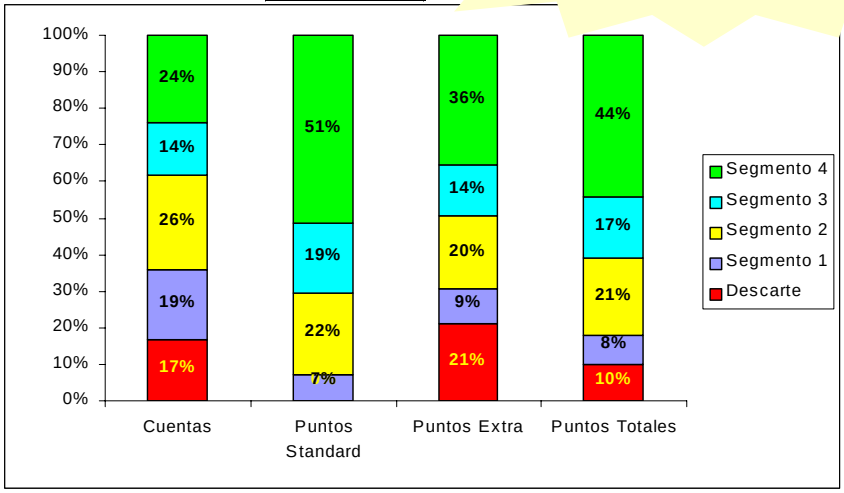
4. Metodología empleada:

- Data cleaning: análisis de valores perdidos (missings) y fuera de rango (outliers).
- Discretización de las variables en tramos por cuartiles para eliminar el sesgo de las distribuciones.
- Homogeneización multivariada
- Cluster K-Medias sobre las dimensiones

Como resultado de la investigación se identificaron 4 segmentos efectivos y 1 de "Descarte".

Segmento Telecom	Cuentas
4-Muy buenos	24%
3-Buenos	14%
2-Regulares	26%
1-Malos	19%
Descarte	17%
	100%

El **Segmento "4-Muy Buenos "** representa el **24% de las Cuentas Telecom-Travelpass** que acumula el **44% de los Puntos Totales**



Internet:

- **El 70% de los Clientes registrados en el site de Travelpass son clientes del Grupo Telecom.**

✓ **El 50% del 70% corresponden al Segmento "4-Muy Buenos" en el Grupo Telecom.**

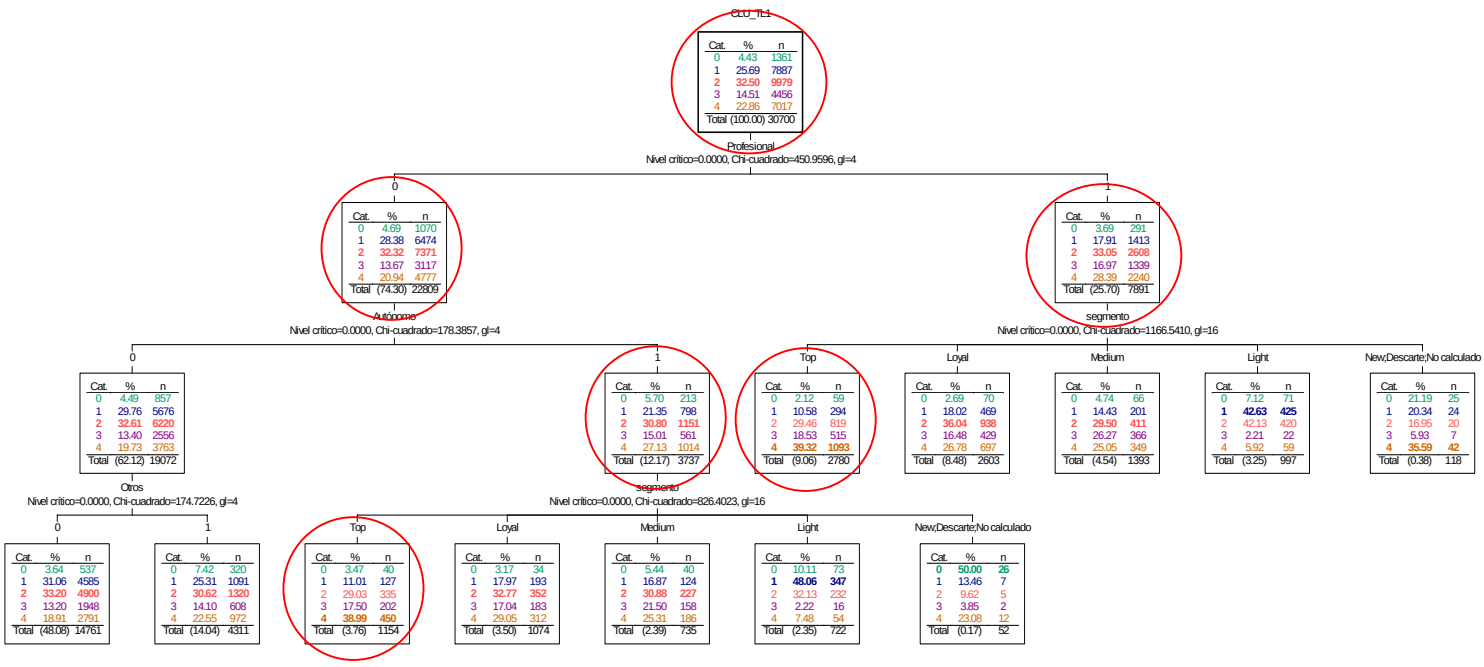
Búsqueda de mellizos:

- La búsqueda de mellizos se planteó caracterizando a los Segmentos Telecom a partir de otras variables disponibles en la base Travelpass: Ocupación (Profesional ó Autónomo) y Segmento Travelpass.

- Para estudiar la relación de estas variables con los segmentos Telecom se utilizó un árbol CHAID.

- El algoritmo identifica automáticamente a los mejores predictores para dicha variable**.**

Árbol de Decisión CHAID – Búsqueda de Mellizos



Conclusiones:

- El Modelo de Segmentación es efectivo para identificar a los mejores clientes del Grupo Telecom en Travelpass.

- Los mejores clientes Telecom se ubican en las categorías de Profesionales y Autónomos - Segmento Top Travelpass.

- La base de datos Travelpass aporta indicadores sociodemográficos y de comportamiento asociados a cada segmento Telecom que permiten realizar acciones personalizadas.

- Los mejores clientes Telecom son proactivos a las comunicaciones a través de Internet.

Ajuste del Modelo de Segmentación Travelpass

El análisis de las variables económicas externas y de contexto y la etapa de maduración y situación interna del Programa, sugirieron un ajuste en el modelo de segmentación inicial.

Si bien la subida de los precios al consumidor (productos de supermercados y combustibles) manifestada durante el primer semestre de 2002, impactó levemente en la Tasa de acumulación de Puntos del Programa, compensado ésto por la receción en el consumo y el cambio de hábitos de los clientes (migración de primeras a segundas marcas) cierto es que las variables de gestión y la etapa de maduración del programa requirieron un **re-enfoque en los clientes más valiosos.**

En Mayo de 2002 se realizó una revisión y ajuste del Modelo de Segmentación. La nueva "corrida del Modelo" se realizó sobre la base conceptual y metodológica del modelo inicial.

El objetivo de este ejercicio fue obtener segmentos que concentraran aún más los clientes más valiosos y activos del Programa.

Además, el nuevo modelo se focalizó en las **Cuentas Activas**, teniendo en cuenta que aquellas no acumulan Puntos por el término de 6 meses son dadas de baja del Programa en forma automática.

El plan de acciones fue el siguiente:

1. Enfocar el universo de estudio a las **cuentas activas** al momento del análisis.
2. Correr el nuevo modelo de segmentación sobre el universo total y analizar las diferencias entre ambos modelos respecto de las variables de segmentación aplicadas: frecuencia, recencia, puntos por mes, multimarca.
3. Analizar en ambas segmentaciones (Modelo Inicial y Nuevo Modelo) la incorporación de nuevas variables como Índice de Redención.
4. Analizar, corregir y decidir los nuevos criterios de segmentación.

Discretización de la variable "Puntos por mes"

Con el objetivo de focalizarse aún más en los clientes más valiosos del Programa, se decidió aumentar los rangos de clasificación para las distintas categorías de la variable "Puntos por mes"

Nueva discretización de la variable:

Variable/Código	Valores para la variable en el Modelo Ajustado **(Nuevo Modelo)**			
	1	**2**	**3**	**4**
Puntos por mes Referencia **P**	150 o más	100-150	50-100	Hasta 50

Valores para la variable en el **Modelo Inicial**			
1	**2**	**3**	**4**
100 o más	50 a 100	20 a 50	Hasta 20

La nueva descripción de la variable "Puntos por mes" para cada uno de los segmentos muestra un incremento en la Tasa de Acumulación de Puntos con el consiguiente "recorte de la base" en niveles más elevados.

En el gráfico siguiente pueden observarse las diferencias de comportamiento:

		P25%	P50%	P75%
Sgto Travelpass	Top	124	162	239
	Loyal	53	64	80
	Medium	26	33	44
	Light	7	11	16
	New	25	45	87
Account status	Activas	59	92	155
	Cerradas	4	16	43

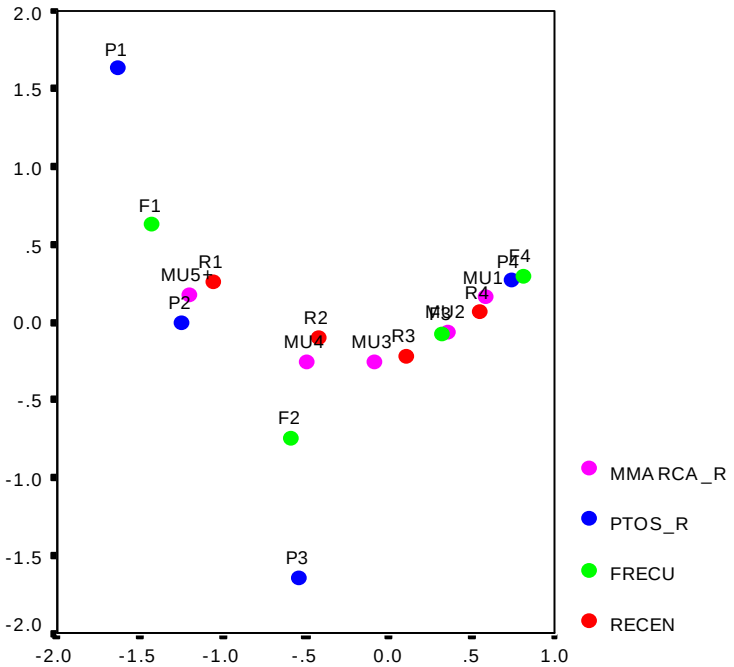
Para el **PLANO COORDENADO DE CATEGORÍAS DE VARIABLES** se consideró a las variables medidas en una escala múltiple nominal (sin ordinalidad).

En el mapa de categorías se observa que la primera dimensión refleja la ordinalidad de las categorías de todas las variables a pesar de que no fue un supuesto de análisis.

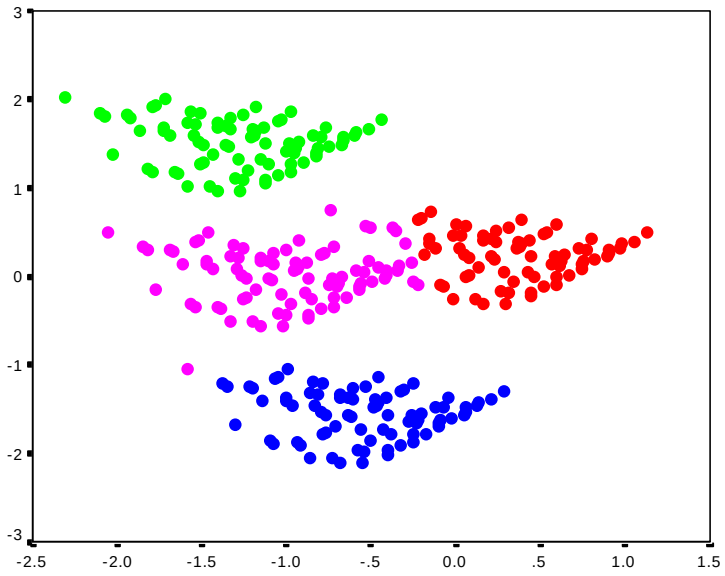
El mapa de individuos muestra la separación de los clientes en los nuevos 4 clusters.

El algoritmo de K-Medias que se aplicó entonces sobre las dimensiones resultantes del análisis en componentes principales condujo a 4 nuevos segmentos (Nuevos Top, Loyal, Medium, Light.) que concentran a los clientes más valiosos del Programa.

Mapa Perceptual : PLANO COORDENADO DE CATEGORÍAS DE VARIABLES
**** NUEVA SEGMENTACIÓN ****
COMPONENTES PRINICIPALES CATEGÓRICAS (CATREG)

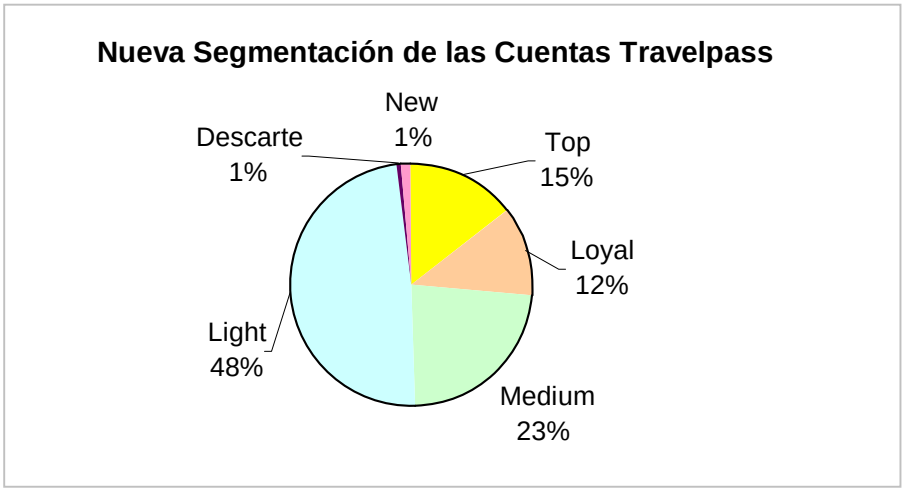


Mapa Perceptual : SEGMENTOS
**** NUEVA SEGMENTACIÓN ****
MAPA DE INDIVIDUOS



Interpretación de los Nuevos Segmentos Travelpass

Para confirmar la claridad de los cluster obtenidos se ha representado el escenario de cuentas y puntos en un diagrama de Pareto. En el diagrama de Pareto puede observarse la participación de estos segmentos en el total de puntos entregados por Travelpass. El 66% de los puntos emitidos se concentra en 2 segmentos (Top y Loyal) que representan el 27% de las cuentas del programa.

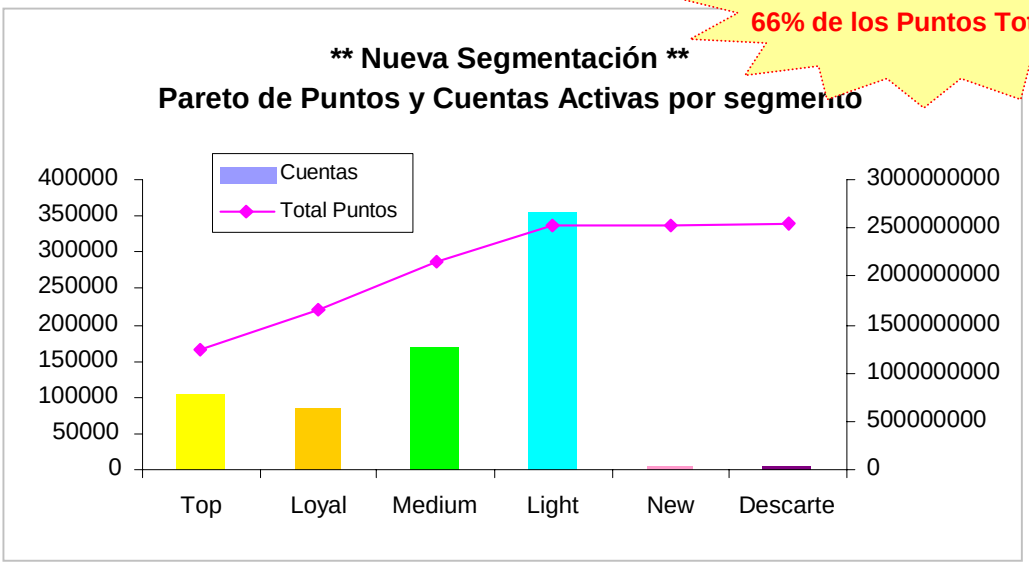


Nueva Segmentación - Cuentas Activas

Nuevo Segmento	% Cuentas		% Puntos Totales	
Top	15%	15%	49%	49%
Loyal	12%	27%	17%	66%
Medium	23%	50%	20%	85%
Light	49%	99%	15%	100%
New	1%	99%	0%	100%
Descarte	1%	100%	0%	100%

Base de Cuentas Activas: **726.853 cuentas**

El 27% de las Cuentas acumula el 66% de los Puntos Totales



A su vez, la descripción de las variables originales para cada uno de los segmentos muestra claramente las diferencias de comportamiento entre los segmentos.

	Variables Descriptivas por Cuartiles			
	Segmentos	**P25%**	**P50%**	**P75%**
Índice Multimarca	Top	3	5	6
	Loyal	3	4	6
	Medium	2	3	4
	Light	1	2	3
	New	1	1	2
Recency	Top	13	19	28
	Loyal	14	19	31
	Medium	18	28	61
	Light	20	35	129
	New	14	23	34
Puntos standard por mes	Top	179	224	314
	Loyal	106	118	132
	Medium	59	69	83
	Light	14	24	35
	New	22	43	82
Frecuencia (dias entre transacciones)	Top	4	6	11
	Loyal	6	8	12
	Medium	8	11	17
	Light	18	25	37
	New	7	14	28
Indice de redencion	Top	0	56%	77%
	Loyal	0	42%	68%
	Medium	0	0%	54%
	Light	0	0%	0%
	New	0	0%	0%
	Descarte	0	0%	0%

	Variables Descriptivas por Nuevos Segmentos				
P50%	**Índice Multimarca**	**Recency**	**Puntos Standard por mes**	**Frecuencia**	**Indice de Redención**
Top	**5**	**19**	**224**	**6**	**56%**
Loyal	**4**	**19**	**118**	**8**	**42%**
Medium	**3**	**28**	**69**	**11**	**0%**
Light	**2**	**35**	**24**	**25**	**0%**
New	**1**	**23**	**43**	**14**	**0%**

Como puede observarse, los segmentos Top y Loyal mantienen similares patrones de uso de la tarjeta, medido en su composición de Frecuencia, Recencia, e índice multimarca, pero con una marcada diferencia en la tasa de acumulación de puntos por mes.

A continuación se muestra el paralelo entre las variables descriptivas de la segmentación Inicial y la Nueva Segmentación para los segmentos Top, Loyal, Médium y Light.

A título de comentario, nótese que en la Segmentación Inicial el Segmento Descarte implicaba el 14% de las cuentas, mientras que en la Nueva Segmentación comprende el 1% de las cuentas. Ello se debe a que en el segundo caso se ha incorporado la variable "Status de la Cuenta = Activa", tomando solamente las cuentas con un Recency menor a 6 meses.

El valor de las variables descriptivas se ve incrementado en distintos los segmentos en el Nuevo Modelo de Segmentación. Ello pone de manifiesto una mayor concentración de mejores clientes en los Nuevos Segmentos. Por otra parte, el hecho de estar trabajando sólo con Cuentas Activas, refina notablemente el proceso de clasificación.

Segmentación Inicial vs. Nueva Segmentación
Variables Categóricas por Segmento

Segmento Top	S_Inicial	S_Nuevo
% Cuentas	**19%**	**15%**
% Puntos	**62%**	**49%**
Indice Multimarca	**4**	**5**
Recency	**18**	**19**
Puntos Standard mes	**166**	**224**
Frecuencia	**7**	**8**
Índice de Redención	No calculado	**56%**

Segmento Loyal	S_Inicial	S_Nuevo
% Cuentas	**16%**	**12%**
% Puntos	**18%**	**17%**
Indice Multimarca	**3**	**4**
Recency	**19**	**19**
Puntos Standard mes	**67**	**118**
Frecuencia	**10**	**8**
Índice de Redención	No calculado	**56%**

Segmento Medium	S_Inicial	S_Nuevo
% Cuentas	**28%**	**23%**
% Puntos	**13%**	**20%**
Indice Multimarca	**2**	**3**
Recency	**68**	**28**
Puntos Standard mes	**35**	**69**
Frecuencia	**22**	**11**
Índice de Redención	No calculado	**56%**

Segmento Medium	S_Inicial	S_Nuevo
% Cuentas	**22%**	**23%**
% Puntos	**4%**	**20**
Indice Multimarca	**2**	**2**
Recency	**143**	**35**
Puntos Standard mes	**11**	**24**
Frecuencia	**43**	**25**
Índice de Redención	No calculado	**56%**

S_Inicial: Segmentación

S_Nuevo: Nueva Segmentación

Determinación del Riesgo de Canje de Recompensas – Un Modelo de Regresión Logísitica

En el marco del Relacionamiento con los Clientes Travelpass se decidió profundizar el estudio de una de las variables más importantes para los Clientes y el Programa: la redención de puntos, el decir el Canje de Recompensas.

Para **los clientes** la obtención de la recompensa en la materialización de la "promesa realizada por Travelpass" y a la que ellos adhirieron.

Para **el Programa**, el índice de redención es una de las variables de gestión más importante que debe enmarcarse en el modelo económico del programa. Los puntos redimidos corresponden un valor crítico del Programa, sea por el nivel de contingencia al que expone así como los recursos necesarios para afrontar las obligaciones con los clientes.

Teniendo en cuenta estas premisas se desarrolló un modelo que permita predecir los clientes que tienen mayor probabilidad de canjear una recompensa y en virtud de ello asignarle un scoring (o valor de probabilidad de canje). Para ello se aplicaron los principios de la técnica conocida como **Regresión Logística**.

La regresión logística es una técnica apropiada cuando la variable dependiente es categórica (nominal o no métrica) y las variables independientes son métricas. Además, la regresión logística predice directamente la probabilidad de un suceso, como expresión de una medida métrica. Los valores de la probabilidad pueden ser cualesquiera entre cero y uno.

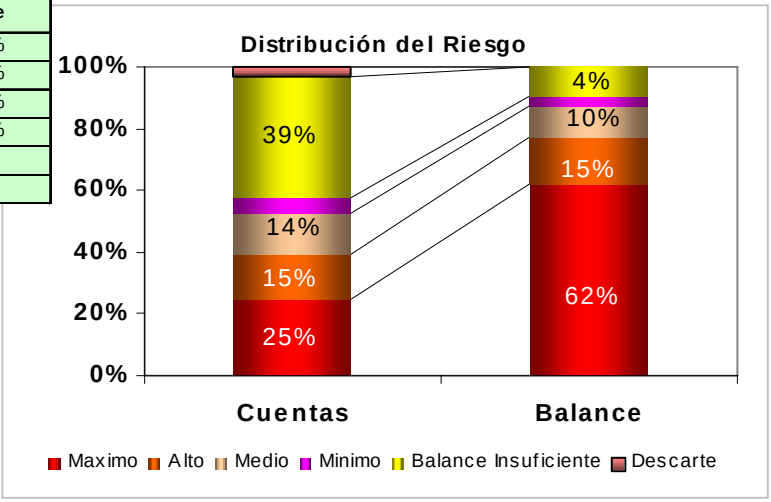
- La variable a explicar es el **Riesgo de Canje**
- El suceso a Predecir es que el cliente "**canjee = 1**" o "**no canjee = 0**"
- Las variables independientes que intervienen en la descripción del Riesgo de Canje son:
    - **Puntos Balance** (balance): saldo de puntos disponibles en la cuenta del cliente al momento del cálculo.
    - **Recency** (recy): recencia en Travelpass calculado como días desde la última transacción al momento del análisis.
    - **Canjes anteriores** (ncanjes): Cantidad de canjes anteriores realizados por el Cliente
    - **Expiración de Puntos**[5] (expir_sn) = 0/1 (No/SI le expiraron puntos)
    - **Recencia desde el último canje** (recy_cje) = días transcurridos desde el último canje
    - **Activación en Travelpass** (activ): Días desde la primera transacción en el Programa Travelpass

Con estas variables se procedió al cálculo de la función Riesgo de Canje, la cual se expresa en términos de las variables independientes descriptivas:

Riesgo = f (balance, acumulación, actividad, canje, expiración)

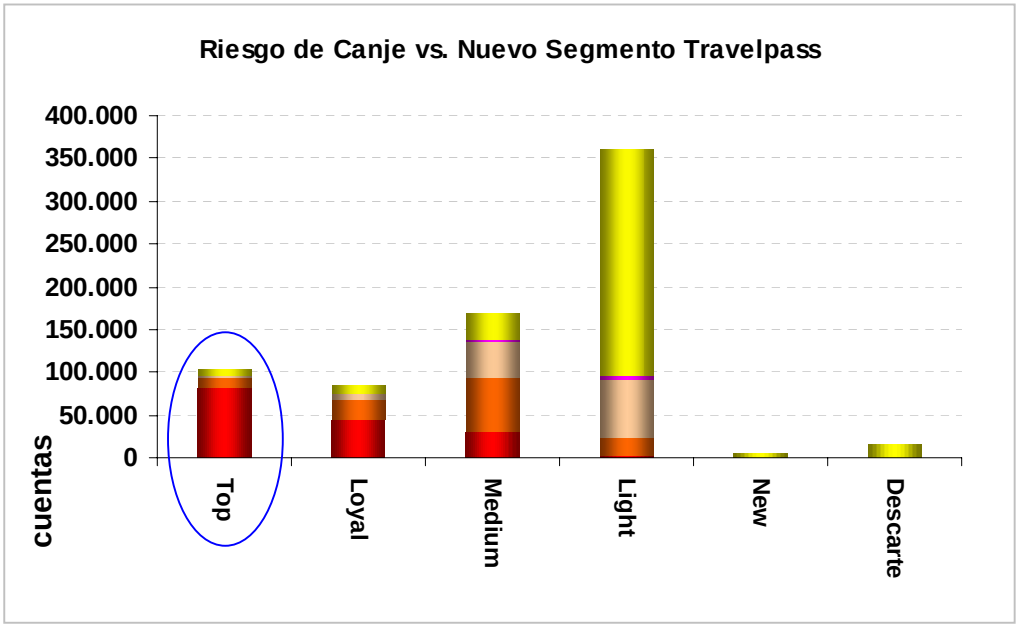
[5] Recuérdese que los Puntos se vencen a los 24 meses de su acreditación.

Riesgo	Scoring	Probabilidad de Canje
Maximo	1	90%
Alto	2	63%
Medio	3	39%
Minimo	4	12%
Balance Insuficiente	5	
Descarte	6	



El 25% de las Cuentas en Riesgo Máximo de canje posee el 62% de los Puntos Balance del Programa.

El 40% de las Cuentas comprendidas en los Niveles de Riesgo Máximo y Alto concentran el 77% de los Puntos Balance del Programa.



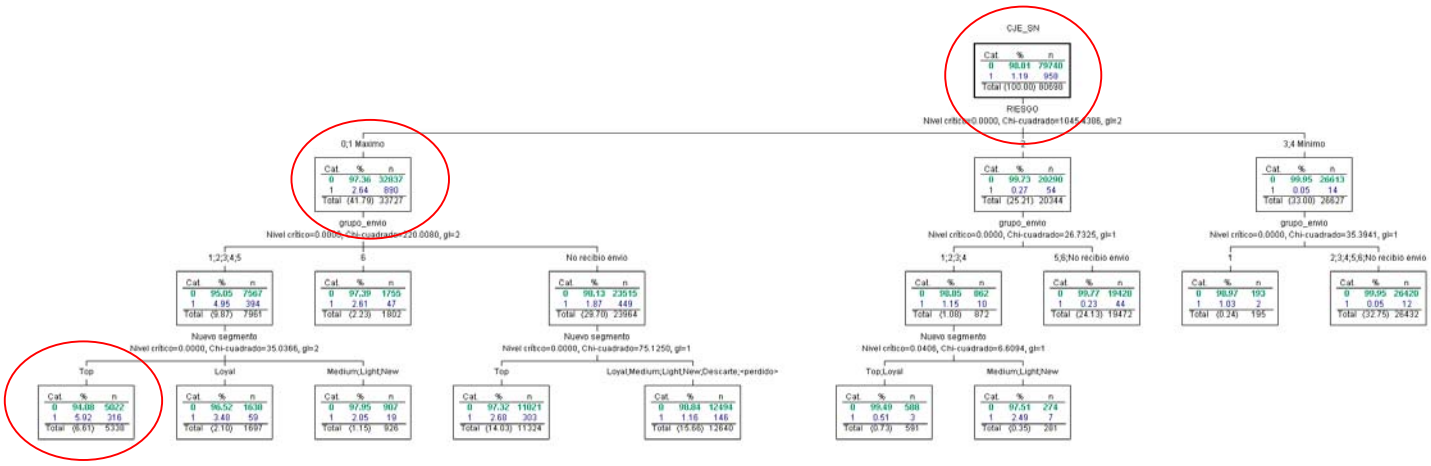
Una vez más se valida la segmentación de Clientes, ya que los clientes más valiosos del Programa son los que poseen una Probabilidad de Canje mayor.

Mediante un Árbol CHAID se analizó el perfil de los clientes que canjearon recompensas con motivo del envío de e-mails realizados durante los meses de Junio y Julio.

El target de los envíos era eran Clientes registrados en la web Travelpass y activos en el Programa.

El objetivo de esta medición fue comprobar cuán ajustado es el modelo predictivo respecto de la situación real.

Árbol CHAID – Medición del Nivel de Canje



Como puede observarse en el Árbol, el 1.2% de los clientes que recibieron los e-mails canjearon recompensas.
Para los clientes asignados bajo el Nivel de Riesgo 1, el porcentaje asciende al 2.64 y si además es un cliente Top (de la Nueva Segmentación) el porcentaje se aproxima al 6%.

Como regla general, la mejor Tasa de Respuesta, medida en Nivel de Canje, se obtiene en los Clientes Top con Nivel de Canje =1 y que reciben Comunicación del Programa Travelpass.

Conociendo las reglas que describen a los clientes más propensos al Canje se puede manejar la variable de Gestión incentivando o retractando la inducción al mismo.

Minería y Análisis de Datos
aplicados en un Programa de Fidelización de Clientes Multimarca

CONCLUSIONES

Como ha quedado demostrado, el Data Mining y las técnicas de Investigación permiten encontrar el verdadero valor de la información, a partir del análisis de los datos transaccionales u operacionales. Esta metodología de trabajo proporciona el ambiente necesario para desarrollar estrategias sobre la base de un conocimiento cierto, analítico, demostrable y hasta metodológicamente "predecibles" de los Clientes de la empresa.

El Programa de Fidelización de Clientes Travelpass es sin lugar a dudas una fuente de gran potencial para desarrollar procesos de investigación ya sea desde la perspectiva del Data Mining como así también de la Investigación de Mercado. Ofrece un calificado "Panel de Consumidores" susceptible de ser analizado desde distintos puntos de vista y enfoques.

Esta integración de metodologías de trabajo –Data Mining e Investigación de Mercado- resulta posible en la medida que (1) se disponga de fuentes de datos accesibles desde el sistema transaccional y (2) exista interés en promover acciones de comunicación con el cliente, que permitan recopilar datos sociodemográficos –hábitos y preferencias, perfil del grupo familiar, ocupación, etc.-. Si a esto se le suma la facilidad, comodidad y economía de contacto de un canal como Internet, el proceso de búsqueda de conocimiento queda expresado en su máxima potencialidad. En este escenario, el Data Mining, la Investigación de Mercado y las tecnologías de la información se integran, potencian y complementan ofreciendo un mayor valor a los resultados obtenidos.

El comportamiento de consumo de los Clientes del Programa, medido a través de sus **variables cuantitativas –*variables duras*-** y que se traduce en los segmentos y otros indicadores, obtenidos de la minería de datos, no necesariamente se corresponde con perfiles sociodemográficos ***–variables blandas-.*** Sin embargo, son estos perfiles de comportamiento los que se necesitan para orientar las Estrategias de Relacionamiento con los Clientes, de la forma que se ha visto expresado en el Caso Travelpass.

Asimismo, es la combinación de perfiles de consumo e información sociodemográfica –obtenida a través de las investigaciones de mercado- lo que permite encontrar el más completo perfil de clientes, ideal para replicar en el resto de la base en un proceso conocido como "búsqueda de mellizos".

Por estos motivos, la integración de la Minería de Datos, la Investigación de Mercado y las técnicas de información, resulta estratégicamente imprescindible para soportar la toma de decisiones. Vincular el perfil de consumo con el perfil sociodemográfico de los clientes es, sin lugar a dudas, potenciar el **"verdadero"** conocimiento sobre los mismos.

Desarrollar el **Ciclo de Relacionamiento de Clientes** sobre la base de un conocimiento metodológico de los mismos, permite optimizar los recursos de la compañía y manejar los indicadores de gestión y performance en un adecuado balance con los niveles de satisfacción de los clientes Travelpass. Ya lo decía Sir Arthur Conan Doyle: **"Es un error capital Teorizar sin tener datos**".

Como ha podido observarse en el desarrollo de la aplicación, los mejores clientes del Programa, representaron el **35% de las Cuentas Travelpass** que acumulaban el **80% de los Puntos Totales**. Esta relación de Cuentas y Puntos –relación de Pareto- evidenciaba claramente la oportunidad de focalizar los recursos del Programa en los clientes más valiosos del Programa, denominados Top y Loyal. Periódicas encuestas de satisfacción han demostrado que éstos clientes son los más satisfechos del Programa, que son receptivos de nuevas propuestas y que mantienen a Travelpass en el "Top of mine" respecto de los demás programas de fidelización. El 98% reconoce a Travelpass como un Programa de Premios por la compra habitual en las empresas, se obtuvo un 95% de Top of Mind en Tarjetas de Fidelización y se alcanzó un Índice de Satisfacción General del 80%, medido en Marzo de 2002.

Los clientes Top y Loyal manifestaban el mejor perfil de comportamiento de consumo, medido a través de sus variables cuantitativas –Índice Multimarca, Recencia, Puntos Standard por mes y Frecuencia en el Programa-. Pero también manifiestan el mejor comportamiento desde el punto de vista actitudinal frente al programa: son los clientes que más valoran los puntos y más recompensas canjean –el 50% de los clientes Top ha canjeado una recompensa-; son los clientes más participativos y receptivos de propuestas de acciones y comunicación y son los clientes que más participación han tenido en Internet (el 87% ha canjeado una recompensa a través de Internet).

Mediante la utilización del Modelo de Segmentación, aplicado a la Estrategia de Comunicación con los Clientes Travelpass y en el marco de una adecuada Estrategia de Premios y Recompensas (asociada al perfil sociodemográfico de los clientes), se logró incrementar en 34% el Índice de Redención de Recompensas (Puntos redimidos/(Puntos entregados), y en tan sólo 10 meses.

La segunda "corrida de modelo" sirvió para enfocarse aún más en los mejores clientes del Programa. Se desarrolló lo que comúnmente se conoce como "descreme" de la base de Clientes. Se ajustaron las variables de acuerdo a las nuevas exigencias económicas del Programa y se obtuvieron segmentos más valiosos desde el punto de vista cuantitativo. Bajo este nuevo esquema, el **27% de las Cuentas Travelpass** acumula **el 66% de los Puntos Totales**. El "nuevo Cliente Top" acumula un 34% más de Puntos Standard Mes que en el modelo anterior así como un "nuevo Cliente Loyal" acumula un 76% de Puntos Standard Mes respecto del modelo inicial.

Como ha podido observarse, una vez que se tuvieron definidos los segmentos de clientes se inició un camino de búsqueda, exploración y aplicación de los mismos, con el objetivo de alcanzar los mejores resultados en todas las acciones realizadas.

Este proceso de **"búsqueda de conocimiento"** tuvo sentido ya que aportó mucho más valor al Programa Travelpass y los resultados han sido superiores a la inversión requerida para obtenerlos.

Además, permitió alcanzar objetivos esenciales para la gestión del Programa Travelpass, tales como:

- **Retención de clientes más valiosos.**
- **Altos Niveles de Satisfacción en los clientes más valiosos.**
- **Optimización de Costos Operativos (atención en el call center; acciones de marketing directo, campañas segmentadas; utilización de Internet como el principal canal de comunicación).**
- **Acciones de Marketing Directo con altas tasas de respuesta.**
- **Sostenimiento de un equilibrado nivel de rentabilidad mediante la administración equilibrada del canje de recompensas** (conociendo los clientes que más propensión tienen al canje y sabiendo que la comunicación es un disparador inmediato a la acción).
- **Mantenimiento del nivel de contingencia del programa, a través de la implementación del Modelo de Riesgo de Canje de Recompensas.**

Paralelamente, se demostró que:

- **La web es un medio eficaz para recabar información sociodemográfica y de perfil de consumo o estilos de vida a un bajo costo por caso.**
- **El análisis metodológico de las respuestas permite describir el "share of wallet" en los mejores clientes y optimizar acciones de retención, recuperación o up selling.**
- **La Minería de Datos es una herramienta poderosa para dar respuesta preguntas complejas de Inteligencia de Negocios.**

Como se ha expresado desde el inicio, el Data Mining no es un fin en sí mismo, ni una solución instantánea a los problemas de negocios, ni un producto que puede comprarse e instalarse cual software "a medida"; sino más bien es un proceso que ayuda a encontrar soluciones a los problemas de negocio y una disciplina que debe dominarse. Podrán aplicarse algoritmos preestablecidos y técnicas milenariamente desarrolladas, pero las cuestiones esenciales quedan aún reservadas al "sentido común" de las personas que "comandan" el desafío de los negocios. El software y la metodología disponibles permiten automatizar parte de la tarea de encontrar los patrones de comportamiento ocultos en los datos, pero hay cuestiones que aún no pueden automatizarse:

- La elección de los problemas de negocio candidatos para tareas de Data Mining.
- La identificación y recolección de los datos que contienen la información buscada y los procedimientos de consolidación y cleaning.
- El "masajeo" y tratamiento de los datos que posibilita la búsqueda de patrones.
- El diseño y cálculo de variables derivadas.
- La selección de los algoritmos apropiados de acuerdo con las características de los datos y las posibilidades del software.
- El análisis crítico posterior de los resultados de los algoritmos de minería.
- El plan de acciones que apoyándose en los resultados del modelo produzca el ROI.
- La medición del éxito de las acciones realizadas a partir de los resultados proporcionados por Data Mining.

Este Proyecto ha servido de gran experiencia para quienes participaron en el desarrollo de las Estrategias como para quienes desean enfrentar el camino del descubrimiento de la información.

Este trabajo ha servido también para demostrar que los grandes proyectos y análisis comienzan por un primer paso y que el principal desafío es la evolución constante y el acompañamiento de la organización, que permita mejorar día tras día. Muchas veces se cree que para explotar estratégicamente los datos de la compañía resulta necesaria una gran inversión que incluya el desarrollo de un mega-proyecto de Data Warehouse previo, antes de poder obtener algún beneficio. Nada más alejado de la realidad que esta conclusión a-priori. Disponer de un Data Warehouse sin duda facilita las tareas de extracción y análisis de datos, pero de ninguna manera es una condición excluyente para **agregar inteligencia a los datos**. Los mejores proyectos de explotación de datos se consiguen a partir de **necesidades puntuales y acotadas,** con resultados medibles en el **corto plazo** y con oportunidades de **re-alimentar la curva de conocimiento de la organización.**

Finalmente y como aportes finales se sugiere tener en cuenta las siguientes consideraciones:

- No es necesario realizar "El Proyecto de Data Mining" para obtener resultados útiles al negocio.
- El desafío no es el proyecto, sino su integración con los objetivos comerciales en el marco de los procesos de decisión.
- La Integración de un equipo multidisciplinarios es fundamental para el logro de los objetivos, generando la "cultura del análisis de datos" a lo largo de todas las áreas de la organización encargadas del desarrollo de estrategias de negocios: marketing, control de gestión, áreas comerciales.
- Antes de seleccionar una potencial aplicación de Data Mining se recomienda considerar:
    - El potencial impacto significativo (relación costo-beneficio).
    - Que exista soporte institucional.
    - Que no existan impedimentos legales de uso y análisis de la información.
    - Que existan suficientes datos.
    - Que existan atributos relevantes.
    - Que los datos posean bajos niveles de ruido.
    - Precisar el nivel de confianza para los resultados.
    - Capacitar a los recursos humanos en las habilidades requeridas para éstos procesos.

Minería y Análisis de Datos aplicados en un Programa de Fidelización de clientes multimarca

REFERENCIAS BIBLIOGRÁFICAS

"Mastering Data Mining -The Art and Science of Customer Relationship Management-", Michael J.A.Berry - Gordon Linoff . John Wiley & Sons, Inc., USA 2000.

"Data Mining Techniques –For Marketing, Sales, and Customer Support-", Michael J. A. Berry - Gordon Linoff. John Wiley & Sons, Inc., USA 1997.

"Data Mining with Neural Networks", Joseph Bigus. Mc Graw Hill, USA 1996.

"Data Preparation for Data Mining", Dorian Pyle. Morgan Kaufmann Publishers Inc., San Francisco, USA 1999.

"Building Data Mining applications for CRM", A. Berson - S. Smith, K. Thearling. Mc Graw Hill, USA 2000.

"Análisis Multivariante" Quinta Edición. Joseph F. Hair, Jr. - Rolph E. Anderson - Ronald L. Tatham - William C. Black. Prentice Hall Iberia, S.R.L., ESPAÑA 1999.

"Una Propuesta Metodológica para la Comunicación de Mapas Perceptuales", V Encuentro de Profesores Universitarios de Marketing - Desarrollo Metodológico de los Profesores Dr. Ramón Pedret Yebra - Asunción Puig Martín – Laura Sagnier Delgado. Sevilla, ESPAÑA 1993.

"20:20 CRM – A visionary insight into unique customer contact", Steve Morrell & Laurent Philonenko. Genesys Telecommunicatrions Laboratories Inc. USA 2001.

"Estrategias de Negocios –Marketing y Estrategia, un enfoque empresarial-", Roberto Sciarroni. Colección coordinada por Carlos Cleri – Mercado. Buenos Aires, ARGENTINA 2001.

"Revista Marketing Directo, Programas de Fidelización: aumentando la lealtad en una época de no compromiso", pág. 14 a 26, Publicación registrada por la Asociación de Marketing Directo, Año 4, Número 19, Buenos Aires, Abril de 2000.

Site en Internet, http://www.pearson-research.com

Site en Internet, http://www.twocrows.com

Site en Internet, The Data Mine http://www.thedatamine.com

Site en Internet, http://kdnuggets.com

Site en Internet, http://www.census.com.ar

Site en Internet, http://www.ibm.com

Site en Internet, http://www.dmg.com

Site en Internet, http://www.ibm.com-redbooks