UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES

A C A T L Á N

"ÁRBOLES DE DECISIÓN, PARA INCREMENTAR LA POBLACIÓN DE LÍDERES DE VENTA EN UNA COMPAÑÍA DE NETWORK MARKETING EN MÉXICO"

T E S I S

QUE PARA OBTENER EL TÍTULO DE

LICENCIADA EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN

P R E S E N T A

MARIA DEL ROCIO NIETO CASAS

ASESORA: M. C. MA. DEL CARMEN VILLAR PATIÑO

Noviembre, 2006

Agradecimientos

- **Agradezco a Dios, la oportunidad de vida para cumplir mis sueños.**

- Gracias a mi Universidad por abrirme las puertas al conocimiento y a la libertad de pensamiento.

- Expreso mi enorme gratitud a la Mtra. Carmen Villar, por su tiempo, asesoría y conocimientos para esta investigación.

- Mi agradecimiento y reconocimiento a los profesores sinodales:

  Mtra. Elvira Beatriz Clavel Díaz

  Mtra. Nora del Consuelo Goris Mayans

  Ing. Ma. Andrea Suárez García

  Lic. Jaime Ramírez Muñoz

por los valiosos conocimientos y aportaciones para elevar la calidad de la tesis.

- A todos mis profesores de MAC, por inculcar amor, dedicación de estudio y disciplina a la carrera.

Dedicatorias

Esta tesis está dedicada a todas aquellas personas que sin su apoyo y confianza no hubiera sido posible realizarla.

A la memoria de mi padre JUAN NIETO, a mi querida mamá PIPINA CASAS y a mí adorada hermana CONCHITA, de los cuales estoy muy orgullosa. Agradezco su ejemplo de lucha y perseverancia en la vida, así como su amor, apoyo incondicional, confianza y fe en todo momento, esta tesis es por y para ustedes.

A mi gran familia, de la cual siempre he recibido ánimo, cariño y motivación, gracias por estar siempre. Ascensión y José; Manuel y Chelo; Ansberto y Socorro; Daniel y Toña; Marce y Lupita; Josefina y Tere; José Luis, Pilar, Luis E., Liliana y Alex; Polo y Vicky; Rafa, Ady y A. Valeria; Tía Dimpna, Martín, Aidé, Rodrigo y Ma. Andrea; Vicky y Dany; Claudia y Manuel; Adela, Coco y Alejandro.

A mis queridas amigas Mary Kay, quienes a través de su experiencia de vida demuestran día con día las grandes mujeres que son, agradezco su gran paciencia, comprensión y apoyo hacia este trabajo: Luz Ma. Almaraz, Graciela Alonso, Mary Guzmán, Eva, Gaby, Alejandra, Carmen y Aída.

A mis queridas amigas y amigos, por la hermosa amistad incondicional, los bellos momentos y experiencias compartidas: Fam. Jacob Cervantes, Gina, Mary, Lucky, América, Elizabeth, Marisol, Sandra, Sofi, Lupita, Nelly, Bertha, Angeles, Marcos, Alfredo, Arturo, Rodolfo, Mauricio, Sergio, Francisco y Eduardo.

Con todo mi amor y admiración,

Rocio

Índice

Índice de figuras

Índice de tablas

Resumen

El estudio tiene como objetivo, encontrar un patrón de comportamiento en la obtención del puesto de líder de ventas entre las personas

distribuidoras que pertenecen a una compañía de venta directa en nuestro país, a partir de la información obtenida al aplicar un cuestionario de reconocimiento empleando la técnica de árboles de clasificación utilizando el algoritmo CART.

El algoritmo se aplica en situaciones donde se tiene un conjunto de datos de individuos en los que se han medido variables predictoras o independientes y una variable de clasificación o de criterio que define el grupo al que cada individuo pertenece, y se quiere encontrar un conjunto de reglas de decisión que permitan explicar la clasificación existente y utilizar estas reglas para poder clasificar un nuevo individuo.

Introducción

"No es que no nos atrevamos porque las cosas son difíciles,
sino que son difíciles precisamente porque no nos atrevemos"

SÉNECA (4 a.C-65 d.C)

Filósofo, dramaturgo y estadista romano

Hoy en día, el sector empresarial para mantener los estándares de calidad que se exigen mundialmente, emplea en sus procesos de trabajo innovadoras herramientas que surgen de los avances científicos en áreas como la tecnología de la información, las telecomunicaciones, el desarrollo de infraestructura entre otras.

En México, a fin de reactivar la economía, instituciones públicas y privadas están impulsando el crecimiento de pequeñas y medianas empresas tanto nacionales como extranjeras, está acción está siendo bien aprovechada por el segmento empresarial logrando incrementar su presencia en el mercado nacional convirtiéndose en líderes en su ramo.

En apoyo a este crecimiento empresarial, este trabajo muestra que el empleo de nuevas herramientas de explotación de información y el adecuado uso de la tecnología, pueden implementarse para optimizar los procesos involucrados en la de toma de decisiones a cualquier nivel de la organización, siempre y cuando se obtengan mayores beneficios en el corto, mediano y largo plazo.

Esta investigación aborda, el caso de una compañía de *network marketing (venta directa)*, que su objetivo *es identificar las características de comportamiento de sus distribuidoras para detectar las fortalezas y áreas de oportunidad de cada una de ellas, con la finalidad de incrementar el número de líderes de venta en un año, reduciendo con ello la inversión de cada reclutamiento, y colocarse dentro de los primeros lugares que ocupan las compañías de venta directa en México.*

Para alcanzar el objetivo, se presenta como alternativa de solución, *el desarrollo de un modelo predictivo de árboles de clasificación*, siendo una técnica de la minería de datos, la cual se considera como una nueva generación de herramientas de análisis de datos para extraer conocimiento útil desde la información que se tiene disponible.

La hipótesis sustentada es que **"el árbol de clasificación generado por el algoritmo CART identifica rápidamente el patrón de comportamiento existente en las distribuidoras que obtienen el liderazgo de ventas en un año, teniendo el poder predictivo para utilizar estas reglas y clasificar eficazmente a grupos futuros de nuevas distribuidoras, con un porcentaje mínimo de error"**.

Sabemos que para explicar *la clasificación*, existen las técnicas tradicionales como el análisis de regresión múltiple, el análisis de regresión logística, el análisis discriminante lineal y cuadrático, etc.

Sin embargo, se eligió la técnica de árboles de clasificación por que encuentra patrones de comportamiento basándose en un conjunto de variables independientes, es de carácter robusto y eficiente para manejar grandes volúmenes de datos, su representación gráfica permite su fácil interpretación y su finalidad es la explicación de la clasificación existente, aprender el patrón de comportamiento y obtener un modelo predictivo para utilizarse en la clasificación de observaciones futuras. Aunado a lo anterior se presenta el algoritmo CART, que es uno de los primeros algoritmos de aprendizaje de árboles y es tanto un clasificador como un árbol de regresión.

El desarrollo de la investigación se presenta como sigue:

Los primeros tres capítulos, presentan el fundamento teórico de la investigación y en los restantes el desarrollo del caso práctico.

El primer capítulo, hace referencia al sistema de venta directa, su origen y evolución, su estructura y esquema de trabajo. También se plantea el problema de la compañía de venta directa, los objetivos a lograr, la hipótesis sustentada y la metodología de trabajo.

El segundo capítulo, se centra en el tema de la minería de datos, de la cual parten los árboles de decisión. Se presentan sus antecedentes, sus fases, los tipos de modelos, las tareas, métodos y algoritmos, así como el software que se encuentra disponible.

El tercer capítulo, abarca a los árboles de decisión, en especial a *los clasificadores,* su definición y estructura, su construcción, procedimientos específicos como la poda y algoritmos propios de árboles.

En el cuarto capítulo, se define la población bajo estudio y se presenta el procedimiento de integración y recopilación de la información, el análisis estadístico descriptivo de la población, el cual sirve para definir las variables predictoras que tienen mayor importancia y las cuales se incluyen en la construcción de los modelos.

El quinto capítulo, aborda la construcción y generación de los árboles clasificadores, utilizando CART y como software el CART versión 5.0. Se especifican los parámetros y variables iniciales que requiere cada modelo.

En el sexto capítulo se evalúan los resultados de los modelos, escogiendo al "mejor modelo predictivo" que obtuvo el algoritmo CART, además se realizan propuestas para la implementación del mismo.

Capítulo 1

CONTEXTUALIZACION DEL PROBLEMA

Palabras Clave: Venta directa. Planteamiento del problema. Objetivos. Hipótesis.

1.1 Conceptos preliminares de la Venta Directa

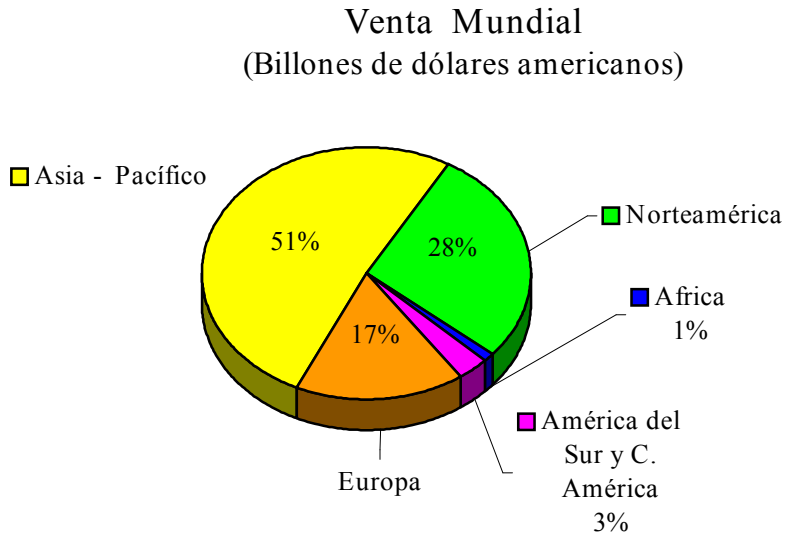
Antecedentes

Las antiguas civilizaciones utilizaban el trueque o intercambio de objetos como una forma de actividad comercial, la cual estaba estrechamente ligada al crecimiento de la economía de sus pueblos. Posteriormente, al implantarse el sistema de intercambio monetario, la base del comercio han sido las ventas de persona a persona, en el pasado artesanos y granjeros que dedicaban su tiempo y esfuerzo a producir primero para su consumo, abandonaban sus trabajos para visitar poblaciones vecinas y vender los artículos producidos por ellos mismos, por esto la venta a domicilio es una de las primeras formas de venta que se conocen y desde este punto de vista las ventas directas podrían tener cientos de años de antigüedad.

Actualmente, la venta directa es básicamente un instrumento de divulgación muy popular a nivel mundial, pero al mismo tiempo complejo, el cual no solamente tiene una gran importancia por las cifras económicas que alcanza, sino por el gran número de fuentes de trabajo que ofrece en todas y cada una de sus facetas empresariales y el impacto que esta industria tiene en las economías de los países es muy fuerte, viéndose reflejado en la transportación, la hotelería, materias primas, componentes, etc. Debido a la importancia económica y social que reviste esta industria, las compañías de venta directa han ganado terreno durante los últimos años, caracterizándose por ofrecer a las personas que participan en ellas una forma de obtener ingresos a la par de un desarrollo personal y profesional en las áreas de ventas y de relaciones públicas, sin sacrificar la libertad de tiempo personal y familiar.

Hoy en día existen alrededor de 2,000 compañías en el mundo y están creciendo a una tasa anual de entre 20 y 30 por ciento, este fenómeno está apoyado en su totalidad por los Foros Económicos Mundiales que sostienen que la mediana y pequeña empresa son el origen y motor del crecimiento de toda economía.

Lo anterior se comprueba por las ventas anuales registradas de 72 billones de dólares americanos que alcanzan 56 países, distribuidos en África, América del Sur y Centroamérica, Europa, Asia-Pacífico y Norteamérica, siendo Asia-Pacífico la mayor generadora de ventas, ver figura 1.1.



Fuente: Memorias del primer Congreso Académico Regional de Ventas Directas.(2000). San José, Costa Rica.

Figura 1.1

1.2 Evolución de la venta directa a nivel mundial

- **Origen y establecimiento**

Las referencias más antiguas del sistema de puerta en puerta, de manera organizada datan del año 1851, cuando el Sr. Isaac Merrit Singer consideró indispensable para vender sus máquinas de coser que tenía que ir al domicilio a demostrar su funcionamiento.

Sin embargo, a mediados de 1860 algunas compañías estadounidenses crearon una división específica de ventas llamada de "persona a persona", iniciando las presentaciones de productos al domicilio con citas previamente elaboradas. De esta

manera se comenzaba a formar oficialmente la industria de la venta directa, que para la primer década del sigo XX la industria había crecido en popularidad.

Las compañías más representativas en este surgimiento fueron: “The Southwestern Company”, “Watkins, Inc”, “Avon Products, Inc” y “The Fuller Brush Company”.

- **Primeros conflictos legales y la auto-regulación**

Durante su evolución, surgieron problemas legales y de aceptación pública de la industria, por lo que la necesidad de autorregularse dio pie a que en 1910, diez compañías de ventas directas estadounidenses se unieron para formar la primer Asociación formal de la industria, denominada Asociación de Ventas Directas o DSA (*Direct Sales Association)*.

Las compañías más representativas en este periodo son: “The Kirby Company”, “Enciclopedia Británica North America” y “Stanhome”, esta última vio surgir de sus filas a algunas de las mujeres que más han influenciado a la industria de las ventas directas y que fundaron compañías de renombre como son Mary Kay Cosmetics, Home Interiors and Gifts, Jafra y Tupperware.

- **Surgimiento de las estructuras de Red y Multinivel**

Para finales de la Segunda Guerra Mundial se suscitó una aceleración en el crecimiento de las compañías de venta directa, trayendo como consecuencia una mayor competencia, y en 1941 se da inicio a las estructuras de red y multinivel.

Las compañías tuvieron que empezar a modificar sus planes de compensación para ofrecer mayores incentivos a los representantes y de esa forma atraerlos o retenerlos. Los cambios incluyeron extender los pagos de comisiones a un tercer nivel de pago, sentando así las bases para los planes con estructura de multinivel.

Otra innovación fue el concepto de rangos de reconocimiento, fundamento de los planes con estructura de red. Ésta etapa, fue sin duda de bonanza para la industria de las ventas directas, las principales compañías en este periodo son: “Tupperware, Inc”, “Regal Ware, Inc”, “Highlights for Children” , “Rich Plan Corporation” , “Jewels by Park Lane”, “Jafra”, “Shaklee Corporation”, “Yves Rocher” , “Amway Corporation”, “Mary Kay Cosmetics, Inc” y “Dudley Products, Inc”.

- **Fortalecimiento de la autorregulación de la industria**

Para 1969, los miembros de la Asociación de Ventas Directas decidieron crear un Código de Ética, el cual pretendía protegerse de resoluciones del gobierno contra la industria, ayudar a los consumidores de fraudes y mejorar la imagen de la industria ante los consumidores. Así mismo, en 1973 se constituye la Fundación para la Educación en Ventas Directas (*Direct Selling Education Foundation DSEF*), dedicada a proporcionar conocimientos y programas académicos sobre los diferentes temas que abarcan las ventas directas. Para finales de los años 70’s la Asociación de Ventas Directas se había establecido en suficientes países como para reconocer que las ventas directas eran una industria global, creándose en 1978 la Federación Mundial de Asociaciones de Ventas.

Las principales compañías durante esta etapa son: “National Safety Asociates, Inc”, “Forever Living Products International” y “Discovery Toys, Inc”.

- **Gran crecimiento de compañías a nivel internacional**

En este período muchas de las compañías de venta directa se expandieron al mercado internacional, lo que originó retos en el ámbito legal y organizacional, percibiéndose la necesidad de fortalecer las asociaciones para la industria.

La mayoría de las compañías que surgieron en este periodo ofrecían servicios como su producto primario o secundario. Previo a esta etapa las asociaciones se enfocaban únicamente a las necesidades de las compañías.

Sin embargo, en este período, se preocupan también por proteger los derechos de los consumidores con respecto a los productos y a las oportunidades de negocio.

También aparecen publicaciones periódicas especializadas en la industria, empresas que ayudan a organizar el lanzamiento de una nueva compañía, consultores que ayudan a los representantes independientes a construir sus organizaciones de ventas e incluso algunos abogados que vieron la necesidad de dedicarse por entero a esta industria. Las compañías representativas de esta etapa son: “Herbalife International, Inc”, “Sunrider International”, “NuSkin International, Inc”, “Melaleuca” y “Money Maker’s Monthy”.

- **Desarrollo de compañías globales de venta directa**

Conforme la industria continuaba adquiriendo credibilidad, comienza atraer a grandes compañías con el potencial de abrir sus puertas internacionalmente desde su inicio. Adicionalmente con la ayuda de *Internet* los esfuerzos de comercialización internacional fueron relativamente sencillos. Aparece *software* especializado en compañías de venta directa y multinivel.

En esta etapa, se observa una tendencia de compañías a gran escala, ya sea adquiriendo a compañías de venta directa o multinivel establecidas o creando divisiones de ventas directas en sus compañías para complementar sus esfuerzos de comercialización. Las empresas representativas en este periodo son: “Nikken”, “Excel Telecomunications” , “The Peoples Network”, “New Vision, International” y “Pre-Paid Legal”.

- **La industria de la venta directa en México**

La venta directa en nuestro país, es una actividad relativamente joven, ya que se estableció formalmente en la década de los años cincuenta, con Stanhome (1955), Avon Cosmetics y lo que hoy es Tupperware (1956). Antes de este período, ya se hacian ventas directas con diversos artículos, pero no de forma oficial. Las primeras compañías enfrentaron el desconocimiento de las autoridades y había sospechas de que no fuera muy legal el sistema de recaudación. Para enfrentar estos retos, se fundó la Asociación Mexicana de Ventas Directas (AMVD), y al mismo tiempo México se convirtió en miembro de la Federación Mundial de Asociaciones de Venta Directa.

Actualmente esta actividad se encuentra sustentada por grandes compañías, la gran mayoría de origen extranjero, y las cuales conforman un sector importante en la economía del país, ya que generan importantes fuentes de trabajo, y sus volúmenes de venta representan crecimientos significativos para diversas industrias.

En la actualidad la AMVD, tiene registradas más de cincuenta compañías dedicadas a la comercialización de productos muy diversos.

1.3 Estructura del sistema de venta directa

Para comprender la forma de operación del sistema de Venta Directa, se introducirán conceptos básicos.

De acuerdo con la Asociación Mexicana de Ventas Directas A.C., **la venta directa** se define como, "la comercialización de bienes de consumo y/o servicios directamente a los consumidores, mediante el contacto personal de un(a) vendedor(a) independiente, generalmente en sus hogares, en el domicilio de otros, en su lugar de trabajo o fuera de un local comercial."[1]

Una compañía de venta directa, "es aquella que como método de mercadotecnia vende sus productos directamente al consumidor final o lo hace a través de representantes independientes, por lo que no presentan tiendas al menudeo abiertas al público en general. Su cadena de distribución es corta, ya que puede llegar a ser directa entre el productor y el consumidor final".[2]

La industria que agrupa a las empresas que utilizan la venta directa como su principal canal de distribución para sus productos o servicios, se dice que utiliza un sistema de *comercialización por redes o también llamado de multinivel o network marketing.*

[1] Asociación Mexicana de Venta Directa. (1997). *Memorias del Primer Seminario Académico.* México D.F.

[2] Laggos, Keith B. (1998). *Direct Sales: an Overview.*United States: Donelley.

Canales de distribución en la venta directa

Un **canal de distribución** *es el medio o conducto que cada empresa elige para la distribución más completa y eficiente de sus productos o servicios al cliente o al consumidor.* Se traduce como el conjunto de intermediarios que utiliza el productor para hacer llegar su producto al consumidor. El intermediario es cualquier persona o institución que está entre el productor y el consumidor final.

La venta directa utiliza dos canales de distribución principalmente, los directos y los indirectos.

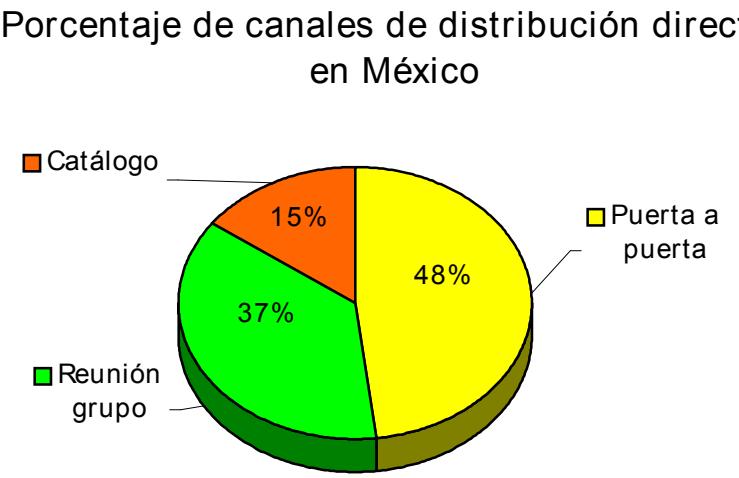
1.- Canales directos. Transfieren la propiedad del productor al consumidor directamente, y los más tradicionales son los siguientes:

a) Reuniones de grupo *(party plan o show room)*. Se realizan reuniones en domicilios particulares, donde la persona anfitriona reúne a un grupo de gente para que el o la representante independiente exponga las bondades de los productos a los consumidores.

b) Puerta a puerta *(door to door).* Las empresas basan su principal fuerza de ventas en personas que se encargan de llevar el producto directamente a la casa o a los centros de trabajo del consumidor.

c) Catálogo *(face to face)*. El o la representante independiente trabaja con un catálogo de ventas para ofrecer los productos ahí descritos.

El porcentaje de la distribución de estos canales en nuestro país se muestra en la figura 1.2.



Fuente: Memorias del primer Congreso Académico Regional de Ventas Directas. (2000).San José, Costa Rica.

Figura 1.2

2.- Canales indirectos. Incorporan a intermediarios comerciantes y se mencionan los siguientes:

a) **Mayoristas.** Son aquellos que venden productos o servicios a las personas que compran con el propósito de revender o con fines industriales.

b) **Minoristas.** Son una empresa comercial que vende bienes o servicios al consumidor final para su uso personal.

La fuerza de ventas

Se encuentra formada por *representantes independientes o también llamados vendedores(as) directos independientes* y son aquellos individuos que participan en representación de sí mismos o de una compañía de venta directa, significa que dichos representantes independientes no son empleados de la compañía que provee los productos que distribuyen, sino personas de negocios independientes que tienen a su cargo la promoción, la entrega y el cobro. Su labor de promoción y ventas es recompensada por un plan de compensación.

Los representantes independientes tienen la oportunidad de obtener ganancias de sus negocios, pero también aceptan la responsabilidad por los riesgos asociados con la operación de los mismos. Los representantes independientes no tienen restricción alguna en cuanto a género, edad, educación ni experiencia previa, ya que las compañías de multinivel, capacitan constantemente a sus distribuidores para la correcta administración de sus negocios o micro-empresas. Para ello ofrecen cursos de capacitación y literatura acerca de como usar y vender los productos, manejo de sistemas de crédito y finanzas, administración de recursos, relaciones humanas, psicología del consumidor, mercadotecnia y publicidad, adecuado uso y manejo de la tecnología, entre otros.

Dentro de su plan de capacitación, enfocan a los distribuidores en una dinámica de logro constante y para ello su principal eje director es la motivación.

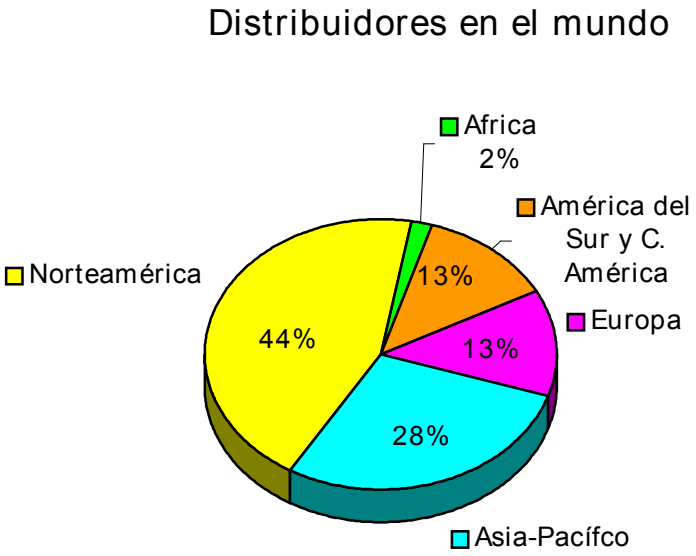
Los distribuidores independientes no se conforman como una empresa típica, sino como una comunidad que comparte un ideal común: la obtención del éxito.

Comulgan ante la idea de que la felicidad y el éxito están en directa correspondencia con la actitud que el individuo tome ante la vida.

Regularmente las personas que se encuentran en los niveles superiores de la red, motivan a los que inician o están en las primeras fases, como se muestra en la figura 1.3.

Figura.1.3 Dinámica de logro constante

A continuación se presenta un estimado del volumen de vendedores o distribuidores independientes en el mundo. Ver Figura 1.4



Fuente: Memorias del primer Congreso Académico Regional de Ventas Directas.(2000). San José, Costa Rica.

Figura 1.4

Un aspecto, que debe destacarse es que el 72% de las personas dedicadas a la venta directa en el mundo son mujeres que van desde los 18 hasta aproximadamente los 65 años de edad. Ver Figura 1.5.

Fuente: Memorias del primer Congreso Académico Regional de Ventas Directas.(2000). San José, Costa Rica.

Figura 1.5

Este aspecto se debe a que el rol de la mujer en la sociedad se ha ido modificando, pasando de ser sólo madre, esposa y ama de casa a ser una parte importante del ingreso familiar, lo cual está fuertemente influenciado por factores culturales y económicos, teniendo que afrontar condiciones de desigualdad, racismo y discriminación. Ver figura 1.6.

Figura 1.6 *La mayoría de las personas dedicadas a la venta directa son mujeres*

Productos y servicios que se comercializan

Las empresas de venta directa ofrecen al público consumidor una gran variedad de productos, entre los que se encuentran los cosméticos y artículos de tocador, perfumería, artículos para el hogar como: electrodomésticos y artículos de limpieza; también ropa, joyería de fantasía, lencería, material educativo, complementos alimenticios y nutricionales, entre otros.

1.4 Esquema de trabajo de la venta directa

La *comercialización multinivel* es un método de organizar y recompensar a los vendedores o distribuidores en un negocio de venta directa. Se define como el *plan de incentivos o de compensación*, mediante el cual los vendedores pueden recibir ingresos de dos formas.

Primero, los vendedores *obtienen descuentos por su volumen personal de ventas de bienes y servicios a los consumidores*, y segundo, *reciben comisiones por las ventas del grupo o red reclutado dentro del plan*.

La compensación en un plan de *comercialización multinivel* se deriva exclusivamente de las ventas de bienes y servicios a consumidores y usuarios finales. Los consumidores finales incluyen a los vendedores que adquieren productos para su uso personal o familiar. No hay ganancia monetaria por el solo hecho de reclutar participantes adicionales en el plan.

Un *plan de compensación*, establece una escala ascendente de retribuciones económicas de acuerdo al nivel de desempeño del vendedor y de su red. Paralelamente al plan de compensación existen otras formas de premiación como bonos en efectivo, viajes o autos de lujo a quien haya cumplido determinado puntaje en un período de tiempo, de esta forma se mantiene motivado al vendedor para elevar su rendimiento al máximo tanto en las ventas propias como en la productividad de su red.

Las características más representativas de la *comercialización multinivel* son:

- El costo inicial de participación es generalmente muy bajo. Normalmente, la única compra requerida es de materiales de capacitación, auxiliares de venta o estuches de demostración.

- Las compañías de comercialización multinivel tienen una fuerte posición en contra de la acumulación de inventario excesivo, dando a los participantes que dejan el plan la oportunidad de devolver cualquier mercancía no utilizada.

- Las compañías de comercialización multinivel son conocidas y respetadas por la calidad de sus productos y la compañía respalda dichos productos con una garantía de satisfacción o derecho de cancelación que permite a los consumidores insatisfechos devolver el producto por un reembolso o crédito comercialmente apropiado.

- Las compañías de comercialización multinivel evitan representaciones de ganancias exageradas para los vendedores que participan en el plan.

1.5 La industria de la Perfumería y Cosmética en México

Como este trabajo estudia el caso de una compañía de venta directa dedicada a la comercialización de cosméticos y artículos de tocador en nuestro país, es necesario hacer referencia a los siguientes antecedentes sobre esta industria en específico.

La industria de la Perfumería y Cosmética agrupa a su mercado en 8 sectores:

1. Productos capilares
2. Tintes
3. Desodorantes
4. Maquillaje y color
5. Perfumes y fragancias
6. Cuidado de la piel
7. Higiene bucal
8. Otros (jabones, gel, cremas y espumas para rasurar, talco, cuidado del sol, bebés, niños, depiladores y varios)

Según la Cámara Nacional de la Industria de la Perfumería y Cosmética (CANIPEC), la venta de perfumería y cosmética mundial ha ido en aumento. Particularmente, en México hubo un crecimiento en la venta de estos artículos del 5.4%, superando a Estados Unidos con un 2.8% y a la mundial de 3.5%.

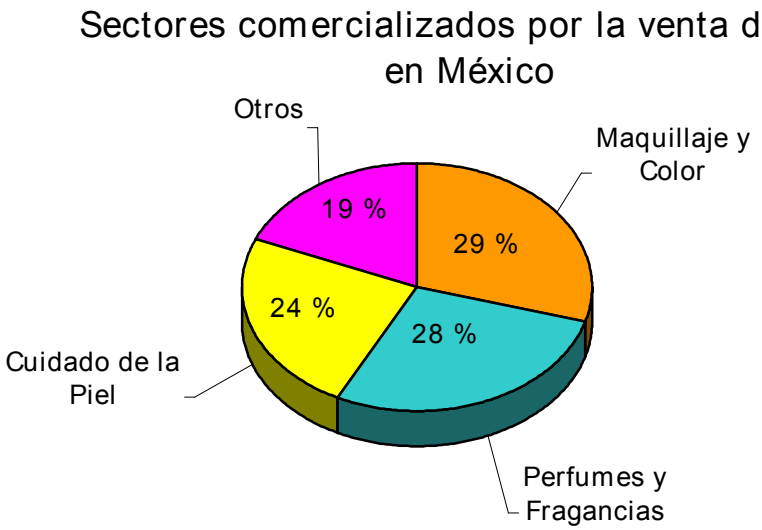
En un estudio que realizó la CANIPEC en nuestro país para analizar los volúmenes de venta de acuerdo a los canales de distribución, resultó que en el año 2003 la venta directa continúa siendo el mayor canal de distribución, comparado con otros sectores como a continuación se muestra en la Tabla 1.1.

Tabla 1.1 Canales de distribución

Canales de venta	Participación (%)
Venta directa	33.40
Autoservicios	31.80
Mayoristas y distribuidores	19.90
Departamentales	5.00
Exportación	3.10
Farmacias	3.00
Gobierno	2.60
Otros	1.20

Fuente: Memoria Estadística.(2000). CANIPEC.

En el año 2004 los principales sectores comercializados por la venta directa en México fueron: maquillajes y color con 29%, perfumes y fragancias con 28% y cuidado de la piel con el 24%. Ver figura 1.7.



Fuente: Memoria Estadística.(2000). CANIPEC.

Figura 1.7

1.5 Planteamiento del problema de la compañía "K" [3], dedicada a la venta directa de Cosméticos y Artículos de tocador en México

Antecedentes

La empresa norteamericana de multinivel K, se dedica a la compra-venta, fabricación y distribución de cosméticos y artículos de tocador a nivel mundial. La empresa tiene alrededor de 30 subsidiarias[4] distribuidas en todo el mundo. Su éxito tanto en los volúmenes de venta como en el número de sus distribuidoras a nivel internacional es muy reconocido y en la actualidad cuenta con más de 1,300,000 personas distribuidoras independientes.

Aunque la subsidiaria mexicana, se encuentra dentro de los primeros lugares a nivel mundial tanto en volúmenes de ventas como en el reclutamiento, otras subsidiarias del continente han venido trabajando arduamente y en cualquier momento pueden desplazar a México de este sitio privilegiado.

Se ha percibido que gran parte de las distribuidoras mexicanas, desempeñan "las ventas" sin mayor problema, alcanzando grandes ganancias económicas y reconocimientos por parte de la compañía, pero con "el reclutamiento" no sucede lo mismo. Sabemos que el multinivel es un negocio de reclutamiento de gente, por lo que las distribuidoras deberán reclutar al mayor número de personas. Una vez estando dentro del negocio deben conservarlos y hacer que escalen los diferentes niveles que presenta el plan de la compañía, para generar las ganancias esperadas.

[3] En esta investigación, por cuestión de confidencialidad la compañía bajo estudio se le denominará como "la compañía K".

[4] Subsidiario-a. (Del lat. subsidiarius). 1.Adj. Que se da o se manda en socorro o subsidio de alguien. 2. Adj.Aplíquese a la acción o responsabilidad que suple o robustece a otra principal.

Planteamiento del problema

En la actualidad, la compañía presenta en su estructura del negocio cuatro niveles básicos que las personas reclutadas pueden ir escalando (Ver Tabla 1.2). El ascenso en los niveles ejecutivos y el porcentaje de retribución se encuentra en función del número de personas que tengan las distribuidoras en sus líneas de patrocinio y de los volúmenes de compra que generen. Se ha detectado que en estos primeros niveles, las personas son más vulnerables de estancarse, o de abandonar la compañía. Al llegar al cuarto nivel y para acceder al nivel denominado *líder de ventas,* la compañía evalúa a las distribuidoras si éstas así lo desean, alcanzando éste nivel las distribuidoras refuerzan su estatus en la compañía, sus porcentajes de compensación se incrementan de manera significativa, además de recibir numerosos reconocimientos y la seguridad de mantenerse en este negocio aumenta. Para la compañía significa crecimiento, estabilidad y mayor presencia en el mercado mexicano.

El tiempo óptimo, para que una distribuidora alcance el nivel de líder de ventas, a partir de ingresar a la compañía, es de un año, ya que la empresa requiere como mínimo cinco meses para que la persona ejerza como distribuidora y empiece a formar su grupo de compra. Una vez cumplido este periodo, la distribuidora podrá solicitar a la compañía que la evalúe y tendrá de uno a cuatro meses para reunir los requisitos que se le solicitan, como son volúmenes de venta, número determinado de afiliadas al grupo de compra, entre otros. Una vez pasando el período de evaluación, la empresa determina si la distribuidora es líder de ventas, de ser negativa la respuesta, la distribuidora podrá repetir el proceso después de seis meses, mientras puede continuar como distribuidora independiente avanzada.

La nueva líder de ventas, tiene como principal objetivo *el desarrollar a nuevas líderes dentro de su grupo de compra*, para seguir creciendo dentro del esquema de la compañía.

Tabla 1.2 Niveles ejecutivos en la compañía K

Niveles		**Puesto ejecutivo**	**Personas reclutadas en el grupo de compra**	**Porcentaje de retribución[5]**	**Beneficios adicionales (seguros de vida y premios de lujo)**	**Líderes desarrolladas en el grupo de compra**
Básico	1	Distribuidora independiente	De 0 a 2	0 %	No	0
	2	Distribuidora independiente con grupo	De 3 a 4	4%	No	0
	3	Distribuidora independiente en formación	De 5 a 7	6%	No	0
	4	Distribuidora independiente avanzada	De 8 a 10	8%	No	0
Avanzados	5	Líder de ventas	De 30 en adelante	12%	Sí	0
	6	Líder de ventas con grupo	Más de 60 personas	12%	Si	1
	7	Líder de ventas en formación	Más de 80 personas	12%	Sí	De 3 a 5
	8	Líder de ventas avanzada	De 100 personas en adelante	12%	Si	De 5 a 8
	9	Líder nacional de ventas	De 300 personas en adelante	12%	Sí	De 8 a 10

[5] El porcentaje de retribución lo evalúa y lo designa la compañía para cada subsidiaria. El porcentaje de retribución presentado en la Tabla 1.2, es el correspondiente a la subsidiaria mexicana.

La compañía ha evaluado que la inversión de recursos que se hace en una nueva persona que ingresa al negocio hasta que llega al liderazgo de ventas es muy elevado, tanto por parte de la compañía como de su líder de grupo. En la actualidad existen personas con más de ocho años en el negocio sin poder ser líderes de ventas, lo cual se traduce en inversión de tiempo, dinero, capacitación y dedicación. La compañía desea que esta inversión se recupere en un año, ya que en otro caso, se manifiestan factores negativos que traen como consecuencia pérdidas en el negocio. Para ello, la compañía se ha dado a la tarea de emprender varias estrategias de trabajo, y los resultados han sido buenos, pero todavía no cumplen con las expectativas planteadas. Entre las estrategias realizadas, la compañía trabaja en conjunto con las líderes de venta exitosas en campañas de motivación, capacitación y concientización entre las distribuidoras que se encuentran en los primeros niveles.

1.7 Objetivos

Los objetivos para la compañía K en México son:

1. Incrementar el número de líderes de venta para que la subsidiaria se posicione en los primeros lugares mundiales.
2. Reducir la inversión de recursos en cada nueva distribuidora.
3. Dar a conocer a mayor número de mujeres el modelo de negocio.

Para colaborar en el logro de los objetivos de la compañía, ésta investigación tiene por objetivo general:

“ Reconocimiento de patrones de comportamiento, que siguen las distribuidoras de la compañía K para la obtención del puesto de líder de ventas, utilizando los árboles de clasificación CART ”.

1.8 Hipótesis

“ El árbol de clasificación generado por el algoritmo CART identifica rápidamente el patrón de comportamiento existente en las distribuidoras que obtienen el liderazgo de ventas en un año, teniendo el poder predictivo para utilizar estas reglas y clasificar eficazmente a grupos futuros de nuevas distribuidoras, con un porcentaje mínimo de error ”.

1.9 Metodología de trabajo

1. Definir la muestra de individuos y las variables que intervienen en el proceso.

2. Recopilación de la información de acuerdo a las variables definidas, mediante un cuestionario de reconocimiento.

3. Realizar un análisis estadístico descriptivo de la información recavada, para conocer el comportamiento de la muestra bajo estudio.

4. Definición de la variable de clasificación o de criterio y las variables predictoras o independientes.

5. Construcción de los modelos de árboles de clasificación utilizando el algoritmo CART.

6. Analizar y evaluar los resultados de los modelos generados para seleccionar el óptimo.

7. Propuestas de cursos de acción sobre el caso de estudio.

Capítulo 2

LA MINERÍA DE DATOS

Palabras Clave: Minería de datos. Tarea. Modelo. Método o técnica. Algoritmos.

2.1 Antecedentes

El concepto de *Minería de Datos (Data Mining)*, surge por la necesidad de encontrar un "nuevo valor" al inmenso volumen y variedad de información que almacenan las empresas en grandes bases de datos y que han ido creciendo exponencialmente. Los datos pasan de ser un "producto" (resultado histórico de los grandes sistemas de información) a ser "materia prima", para explotar y obtener el verdadero "producto elaborado", *el conocimiento*, que es valioso para la ayuda en la toma de decisiones.

La mayoría de la información es histórica y sirve para explicar el pasado, entender el presente, y predecir los eventos futuros. Para analizar los datos, y obtener las explicaciones y predicciones existen diversas herramientas estadísticas tradicionales, pero algunos problemas y limitaciones para trabajar con cientos de tablas, millones de registros de varios *gigabytes* aunado a las nuevas características de los datos como

son los espaciales[6], temporales[7], textuales o documentales[8], multimedia[9] y atributos nominales con muchos valores, han hecho surgir una nueva generación de herramientas y técnicas para extraer de conocimiento útil desde la información que se tiene disponible.

La actual tecnología de *Internet* y su constante crecimiento necesita del desarrollo de tecnologías más avanzadas de minería de datos para interpretar adecuadamente la información y el conocimiento que se encuentra en los datos distribuidos alrededor del mundo.

De esta manera surge la minería *web, que* se define como "el uso de técnicas de minería de datos (Text Mining) para descubrir y extraer información automáticamente desde el World Wide Web"*.*[10] La enorme cantidad de referencias recogidas durante una búsqueda en Internet ilustra muy bien este problema.

La minería de datos, representa *el conocimiento* a través de relaciones, patrones o conjuntos de reglas, conocidas como *ecuaciones, árboles de decisión, redes neuronales, grafos probabilísticos, etc.*

Por el amplio espectro de problemas que encuentran soluciones factibles y certeras en estas herramientas, cada vez es más frecuente encontrar áreas donde se aplica y, algunos ejemplos son:

---

[6] Incluyen datos geográficos, imágenes médicas, redes de transporte, información de tráfico, etc.

[7] Almacenan datos que incluyen una gran cantidad de atributos relacionados con el tiempo, es decir, pueden referirse a distintos instantes o intervalos temporales .

[8] Estas bases de datos pueden contener documentos no estructurados (como una biblioteca digital de novelas), semi-estructurados, (se puede extraer la información por partes, con índices, etc) o estructurados (como una base de datos de fichas bibliográficas).

[9] Soportan objetos de gran tamaño, ya que almacenan imágenes, vídeo y audio.

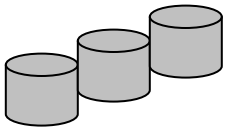
[10] Etzioni, O. (1996) *The World Wide Web:quagmire or gold mine?*. Communications of the ACM.

- **Análisis de mercado.** Detección de los hábitos de compra en los supermercados (también conocido como el análisis de la Cesta de Compra), evaluación de campañas publicitarias, segmentación de clientes, análisis de fidelidad de clientes, estimación de inventarios, costos y ventas.
- **Finanzas y banca.** Obtención de patrones de uso fraudulento de tarjetas de crédito, determinar el gasto de tarjeta de crédito por grupos, análisis de riesgo en créditos, identificación de reglas de mercado a partir de históricos.
- **Medicina.** Identificación de patologías, diagnóstico de enfermedades, predicciones para el mejor uso de recursos, consultas, salas y habitaciones.
- **Telecomunicaciones.** Establecimiento de patrones de llamadas, modelos de carga en redes, entre otros.
- **Web y correo electrónico.** Análisis del comportamiento de los usuarios, detección de fraude en el comercio electrónico, análisis de los logs de un servidor web, clasificación y distribución automática de correo, detección de correo spam, gestión de avisos, análisis del empleo de tiempo, etc.

2.2 El Proceso de Extracción de Conocimiento y sus Fases

La minería de datos, a su vez forma parte del **proceso de extracción de conocimiento también conocido como *Knowledge Discovery in Databases* (KDD**), que se define como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos." [11] Ver figura 2.1.

[11] Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth,P. (1996). *The KDD Process For Extracting Useful Knowledge from Volumes of Data.* Communications of the ACM.



Gran volumen de información histórica

Extracción de conocimiento

Conocimiento nuevo, ayuda en la toma de decisiones

Figura 2.1 *Proceso de extracción de conocimiento*

Elementos de la definición:

1. Patrones válidos.- Los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.

2. Novedosos.- Deben aportar algo desconocido para el sistema y preferiblemente para el usuario.

3. Potencialmente útiles.- Los patrones de la información deben conducir a acciones que reporten algún tipo de beneficio para el usuario.

4. Comprensibles.- La extracción de patrones no comprensibles dificulta su interpretación, revisión, vali

dación y uso en la toma de decisiones.

Descripción de las fases del *KDD*

El proceso de extracción de conocimiento (KDD), se encuentra distribuido en cinco fases las cuales son:

1ª. Fase: Integración y recopilación de información. Aquí se determinan las fuentes de información que pueden ser útiles, se recavan los datos necesarios y se transforman a un formato común, depositándolos en almacenes de datos conocidos como *data-warehouses,* en los cuales se unifica toda la información recolectada. Los almacenes de datos no son estrictamente necesarios para realizar minería de datos, aunque sí extremadamente útiles si se va a trabajar con grandes volúmenes de datos, que varían con el tiempo y dónde se desea realizar tareas de minerías de datos variadas, abiertas y cambiantes.

Este almacén facilita la navegación y visualización previa de sus datos, para ver qué aspectos pueden interesar para ser estudiados.

Las herramientas más utilizadas en esta fase son el procesamiento analítico en línea (O*n-Line Analytical Processing,* **OLAP**)[12] y el procesamiento transaccional en línea (*On-Line Transaction Processing,* **OLTP**)[13].

[12] El procesamiento analítico en tiempo real engloba un conjunto de operaciones, exclusivamente de consulta, en las que se agrega y cruza gran cantidad de información. El objetivo de estas consultas es realizar informes para el apoyo en la toma de decisiones.

[13] El procesamiento transaccional en tiempo real constituye el trabajo primario en un sistema de información. Consiste en realizar transacciones (actualizaciones y consultas) a la base de datos con un objetivo operacional: hacer funcionar las aplicaciones de la organización. Es un trabajo diario y para el que inicialmente se ha diseñado la base de datos.

2ª. Fase: Selección, limpieza y transformación de los datos. La calidad del conocimiento no sólo depende del algoritmo de minería utilizado, sino también de la calidad de los datos, por lo que la preparación de estos datos tiene como objetivo *la eliminación del mayor número de datos erróneos o inconsistentes e irrelevantes*, para presentarlos de la manera más apropiada a la minería.

3ª. Fase: Minería de datos. Esta fase es la más característica del proceso KDD, y su objetivo principal es *encontrar modelos o patrones inteligibles a partir de los datos,* para que sea efectivo, el proceso debería ser automático o semi-automático y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, beneficios a la organización. Para ello es necesario tomar una serie de decisiones antes de empezar el proceso:

- **Determinar qué tipo de *tarea* de minería es la más apropiada.** Una tarea es un tipo de problema a ser resuelto por un algoritmo de minería de datos. Por ejemplo, se podría utilizar *la clasificación* para predecir en una entidad bancaria los clientes que dejarán de serlo.
- **Elegir el tipo de *modelo*.** Los tipos de modelos en minería de datos son predictivos y descriptivos. Por ejemplo, para una tarea de clasificación (predictiva) se utilizaría un árbol de decisión, porque se quiere obtener un modelo en forma de reglas.
- **Elegir *el algoritmo*.** El algoritmo resuelve la tarea y obtiene el tipo de modelo que estamos buscando. Por ejemplo, para la creación de árboles de decisión para clasificación se puede utilizar CART o C5.0, entre otros.

La minería de datos mantiene estrecha relación con otras áreas, entre las que se encuentran (figura 2.2):

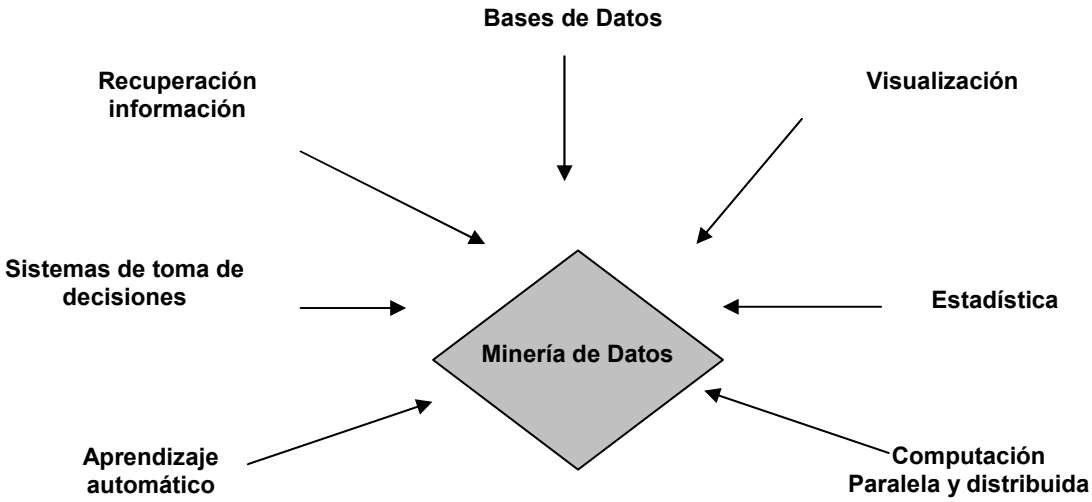


Figura 2.2 *Áreas que contribuyen con la minería de datos*

- *Las bases de datos.* Conceptos como los almacenes de datos y el procesamiento analítico en línea (OLAP) tienen una gran relación, así como las técnicas de indización[14] y de acceso eficiente a los datos son muy relevantes para el diseño de algoritmos eficientes de minería de datos

- *La recuperación de la información*. Consiste en obtener información desde datos textuales, por lo que su desarrollo histórico se ha basado en el uso efectivo de bibliotecas (digitales) y en la búsqueda por Internet.

[14] Es el término correcto para "*indexing*", aunque con frecuencia se utiliza el término "indexación".(Hernández,2004).

- *El Aprendizaje automático (Machine Learning[15]).* Es el área de la inteligencia artificial que se ocupa de desarrollar métodos computacionales para los procesos de aprendizaje[16] y la aplicación de los sistemas informáticos de aprendizaje en problemas prácticos. Los principios seguidos en el aprendizaje automático y en la minería de datos son los mismos: "la máquina aprende un modelo a partir de *ejemplos* o casos y lo usa para resolver el problema".

- *Los sistemas para la toma de decisión*. Son herramientas y sistemas que tienen por objetivo proporcionar la información necesaria para realizar decisiones efectivas en el ámbito empresarial.

- *La visualización de los datos.* El uso de técnicas de visualización, como son las gráficas, las icónicas (basadas en figuras), las jerárquicas (dividiendo el área de representación en regiones dependiendo de los datos), permiten al usuario descubrir, intuir o entender patrones que serían más difíciles de "ver" a partir de descripciones matemáticas o textuales de los datos.

- *La computación paralela y distribuida.* Actualmente, muchos sistemas de bases de datos comerciales incluyen tecnologías de procesamiento paralelo y distribuido. En estos sistemas, el costo computacional de las tareas más complejas de minería de datos se reparte entre diferentes procesadores. Su éxito se debe en parte a la explosión de los almacenes de datos (su adaptación distribuida) y de la minería de datos, en los que las prestaciones de los algoritmos de consulta son críticas. Una de las principales ventajas del procesamiento paralelo es precisamente la escalabilidad de los algoritmos, lo que lo hace idóneo para estas aplicaciones.

[15] Machine Learning. Es un sub-área de Inteligencia Artificial que está dentro del campo de ciencias de computación.

[16] Aprendizaje. Cambios adaptativos en el sistema para hacer la misma tarea(s) de la misma población de una manera más eficiente y efectiva la próxima vez.

- *La Estadística.* Esta disciplina se considera “la madre” de la minería de datos, ya que muchos conceptos, algoritmos y técnicas se toman de ella, pero también presentan grandes contrastes y diferencias, que las hacen dos áreas del conocimiento distintas y no a la minería de datos, como un simple subconjunto de la estadística. Entre las principales diferencias de la minería de datos y la estadística se encuentran:

1) La minería de datos es un proceso que invierte la dinámica del método científico de la siguiente manera, en el método científico, primero se formulan las hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten las hipótesis. Si esto se hace con la formalidad adecuada, se obtiene un nuevo conocimiento. En la minería de datos, se coleccionan los datos y esperamos que de ellos emerjan hipótesis.

2) Las técnicas de minería de datos construyen el modelo de manera automática mientras que las técnicas estadísticas “clásicas” necesitan ser trabajadas por un estadístico profesional.

3) Las técnicas estadísticas se centran generalmente en técnicas confirmatorias, mientras que las técnicas de minería de datos son generalmente exploratorias.

4) A mayor dimensionalidad del problema, la minería de datos ofrece mejores soluciones. Cuantas más variables entran en el problema, más difícil resulta encontrar hipótesis de partida interesantes.

5) El objetivo de la investigación estadística es encontrar causalidad. Si se pretende determinar cuáles son las causas de ciertos efectos (por ejemplo, si invertir más en la publicidad de cierto producto tiene como consecuencia un incremento de ventas o si es más determinante el ofrecer un descuento a los clientes), deberemos utilizar técnicas de estadística (por ejemplo, ecuaciones estructurales). Las relaciones complejas que subyacen a técnicas de minería de datos impiden una interpretación certera de diagramas causa-efecto.

6) Si se pretende generalizar sobre poblaciones desconocidas en su globalidad y si las conclusiones han de ser extensibles a otros elementos de poblaciones similares habrán de utilizarse técnicas de inferencia estadística. Esto viene relacionado con situaciones en las que se dispone exclusivamente de muestras (con el consiguiente problema de aportar validez a las muestras).

7) En minería de datos, se generarán modelos y luego habrán de validarse con otros casos conocidos de la población, utilizando como significación el ajuste de la predicción sobre una población conocida (es lo habitual cuando queremos predecir perfiles de clientes, que ya disponemos de antecedentes para poder validarlos, aunque no siempre es posible acceder a dicha información o no siempre es correcto aplicar ciertas muestras).

Se han comentado algunos argumentos acerca de cuándo es conveniente utilizar minería de datos o estadística, pero debe destacarse que no son excluyentes una de la otra

Así pues, la minería de datos y la estadística son técnicas complementarias que permiten obtener conocimiento inédito en nuestros almacenes de datos o dar respuestas a cuestiones concretas del negocio.

Dependiendo del tipo de datos a ser minados o del tipo de aplicación, la minería de datos usa también técnicas de otras áreas como el lenguaje natural, el análisis de imágenes, el procesamiento de señales, los gráficos por computadora, etc.

4ª. Fase: Evaluación e interpretación de resultados. Medir la calidad de los patrones descubiertos por un algoritmo de minería no es un problema sencillo, ya que esta medida depende de varios criterios subjetivos. Idealmente, los patrones descubiertos deben tener tres cualidades: *precisos, comprensibles e interesantes.* Según las aplicaciones puede interesar mejorar algún criterio y sacrificar ligeramente otro. Para evaluar un modelo, primeramente debe entrenarse y probarse, por lo se dividen los datos en dos conjuntos: un conjunto será el de entrenamiento *(training set)* y el otro será el de prueba o validación *(test set).* Esta separación es necesaria para garantizar que la validación de la precisión del modelo es una medida independiente. Si no se utilizan diferentes conjuntos de entrenamiento y de prueba, la precisión del modelo será sobreestimada y se obtendrán estimaciones muy optimistas, lo cual puede estar muy alejado de la realidad. La precisión en estos conjuntos de datos es una buena referencia de cómo se comportará el modelo para conjuntos de datos futuros. Cada una de las tareas así como cada algoritmo de minería de datos, tiene su propia técnica de evaluación.

5ª. Fase: Difusión, utilización y monitoreo. Una vez validado, el modelo puede ser usado para recomendar acciones basándose en el modelo y en sus resultados, o bien para aplicar el modelo a diferentes conjuntos de datos. Tanto en el caso de una aplicación manual o automática, es necesario su difusión entre los posibles usuarios, e igualmente importante es medir su evolución del modelo, principalmente porque los patrones pueden cambiar, por lo que el modelo debe ser monitoreado, para ser re-evaluado, re-entrenado y posiblemente reconstruido completamente cuando sea necesario.

En la figura 2.3, se presenta el diagrama de flujo de las fases del proceso de descubrimiento de conocimiento (KDD).

El KDD, es un proceso *iterativo e interactivo,* ya que la salida de alguna de las fases puede hacer volver a pasos anteriores y con frecuencia son necesarias varias iteraciones para extraer conocimiento de alta calidad. También es *interactivo* porque el usuario interactúa con un experto para ayudar en cada una de las fases del proceso.

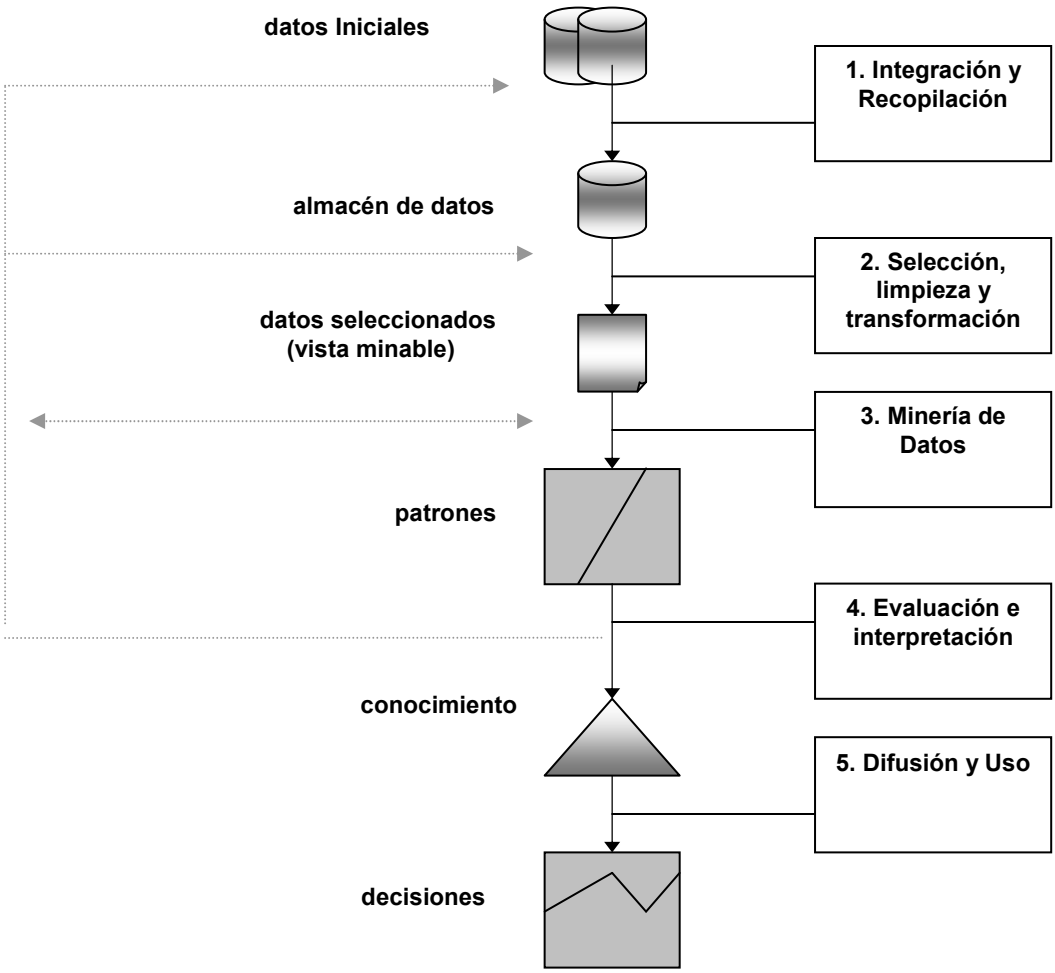


Figura 2.3 *Fases del proceso de descubrimiento de conocimiento (KDD)*

Cabe mencionar que antes de diseñar un plan KDD, debe establecerse de forma clara y concisa el contexto del negocio, los problemas y los objetivos del mismo, con la finalidad de medir la factibilidad de llevar a cabo tareas de minería de datos. Es muy importante la correspondencia entre los objetivos de negocio y los de minería de datos. Ver figura 2.4.

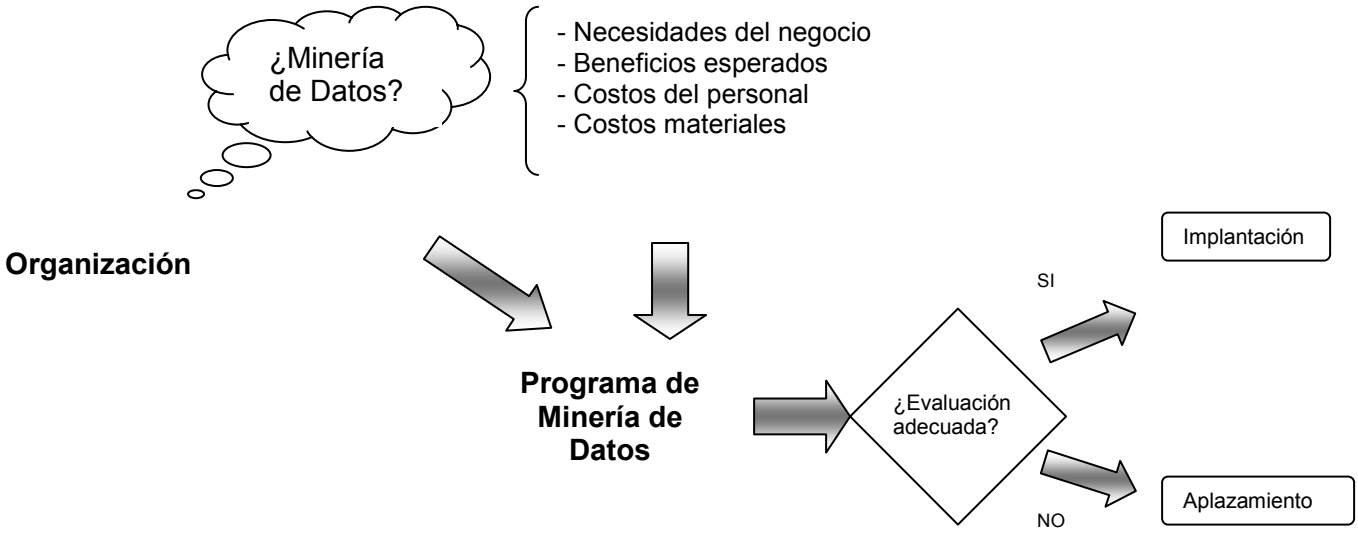


Figura 2.4 *Correspondencia entre los objetivos del negocio y los de minería de datos*

Debe evaluarse los beneficios contra los costos, ya que al implementar minería de datos se puede requerir una inversión considerable en formación, herramientas y personal, aún así la minería de datos debe proporcionar más beneficios que costos, en otro caso deben considerarse otras soluciones o replantearse los objetivos iniciales. Una vez aprobado el proceso de minería de datos, debe seguirse con la metodología adecuada para su desarrollo.

2.3 Tareas, Modelos, Métodos y Algoritmos

De acuerdo al tipo de problema a ser resuelto, en la fase de minería de datos se determina *la tarea* que es más apropiada utilizar, así como *el modelo* y *el algoritmo* (método) más adecuado. Una tarea puede tener algoritmos diferentes para su resolución, y el mismo algoritmo puede resolver un gran espectro de tareas, esto se debe a que la mayoría de las tareas pertenecen *al aprendizaje inductivo o supervisado*, el cual, es un tipo especial de aprendizaje que parte de casos particulares (de aquí en adelante los llamaremos *ejemplos*) y obtiene casos generales (reglas o modelos) que generalizan o abstraen la evidencia. El aprendizaje puede ser además de varios tipos (incremental, interactivo, etc) según ciertas características de la presentación de los datos y de la forma de aprender. Al final del capítulo en la figura 2.6, se presenta un esquema general de la relación que guardan estos conceptos.

2.3.1 Tipos de tareas en minería de datos

Se distinguen dos tipos, *las predictivas y las descriptivas;* cada una de las cuales se considera un tipo de problema a ser resuelto por algoritmos específicos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra.

Definición

Para definir las tareas, primero se define el conjunto de *ejemplos* (casos particulares) con los que se van a tratar. Se define E como el conjunto de todos los posibles elementos de entrada.

Las instancias posibles dentro de E generalmente se representan como un conjunto de valores para una serie de atributos (nominales o numéricos)[17]. Es decir $E = A1xA2x...xAn$ y un *ejemplo* e es una tupla $\langle a1, a2, ..., an \rangle$ tal que $aj \in Ai$ .Ver Figura 2.5

$$E = A1xA2x...xAn$$

$E =$

# instancia	Pronóstico	Humedad	Viento	jugar
1	Soleado	Alta	Débil	No
2	Cubierto	Alta	Débil	Sí
3	Lluvioso	Alta	Débil	Sí
4	Lluvioso	normal	Fuerte	No
5	soleado	normal	Débil	Sí
....	....	....	....	....

← **atributos**

← ***ejemplos***
tupla
$e = \langle a1, a2, ..., an \rangle$

Figura 2.5 Conjunto de ejemplos en una base de datos

Tareas predictivas

Tratan problemas en los que hay que predecir valores para uno o más *ejemplos*. Los *ejemplos* van acompañados de una salida (*clase, categoría o valor numérico*) o un orden entre ellos. Dependiendo de la correspondencia entre los *ejemplos* y los valores de salida y la presentación de los *ejemplos*, se definen varias tareas predictivas tales como, *la clasificación y sus variaciones, la categorización, preferencias o priorización y la regresión,* siendo *la clasificación,* la más utilizada en minería de datos.

[17] Los atributos numéricos contienen valores enteros o reales. Por ejemplo, el salario o la edad.

Los atributos categóricos o nominales toman valores en un conjunto finito y preestablecido de categorías. Por ejemplo, el sexo (H,M).

a) ***Clasificación.*** Cada instancia o registro de la base de datos pertenece *a una clase*, la cual se indica mediante el valor de un atributo que llamamos *la clase de la instancia.* Este atributo puede tomar diferentes valores discretos, cada uno de los cuales corresponde a una clase. El resto de los atributos de la instancia (los relevantes a la clase) se utilizan para predecir la clase. *El objetivo es predecir la clase de nuevas instancias de las que se desconoce la clase.* El algoritmo maximiza la razón de precisión de la clasificación de las nuevas instancias, la cual se calcula como el cociente entre las prediciciones correctas y el número total de predicciones (correctas e incorrectas).

En la clasificación, los *ejemplos* se presentan como un conjunto de pares de elementos de dos conjuntos, $\delta = \{\langle e,s\rangle : e \in E, s \in S\}$, donde S es el conjunto de valores de salida.

Los *ejemplos* e, al ir acompañados de un valor de S, se denominan comúnmente *ejemplos* etiquetados $\langle e,s\rangle$ y, en consecuencia, δ se denomina conjunto de datos etiquetado.

El objetivo es aprender[18] una función $\lambda : E \to S$, denominada *clasificador,* que represente la correspondencia existente en los *ejemplos*, es decir, para cada valor de E tenemos un único valor para S.

Además, S es nominal, es decir, puede tomar un conjunto de valores $c1, c2, ..., cm$ denominados clases (cuando el número de clases es dos, tenemos lo que se llama *clasificación binaria*). La función aprendida será capaz de determinar *la clase* para cada nuevo *ejemplo* sin etiquetar, es decir dará un valor de S para cada valor de e.

[18] Los principios seguidos en el aprendizaje automático y en la minería de datos son los mismos: "la máquina aprende un modelo a partir de ejemplos y lo usa para resolver el problema".

Ejemplo. Consideremos a un oftalmólogo que desea disponer de un sistema que le sirva para determinar la conveniencia o no de recomendar la cirugía ocular a sus pacientes. Para ello dispone de una base de datos de sus antiguos pacientes clasificados en operados satisfactoriamente o no en función del tipo de problema que padecían (miopía y su grado, o astigmatismo) y su edad. El modelo encontrado se utiliza para clasificar nuevos pacientes, es decir, para decidir si es conveniente operarlos o no.

Una variante de *la clasificación,* se denomina ***la clasificación suave***, donde también tenemos pares de elementos de dos conjuntos, $\delta = \{\langle e,s\rangle : e \in E, s \in S\}$ y además de la función λ se aprende otra función, llamada $\Theta : E \to R$, que significa el *grado de certeza de la predicción* hecha por la función λ. Se prefiere tener un clasificador suave con *una medida de certeza* que acompañe a las predicciones. Este tipo de extensión permite realizar otras aplicaciones como son los rankins[19] de predicciones o la selección de los *n mejores ejemplos*.

Ejemplo. Se clasifica un mensaje de correo electrónico como spam o no, proporcionando además la certeza de la clasificación.

La estimación de probabilidad de clasificación, es una generalización de la anterior. La presentación del problema es igual que *la clasificación normal y suave,* pares de elementos de dos conjuntos, $\delta = \{\langle e,s\rangle : e \in E, s \in S\}$.

[19] Ranking. (Voz ingl). 1. m. Clasificación de mayor a menor, útil para establecer criterios de valoración.

Sin embargo, la función a aprender es distinta, ya que se trata de aprender exclusivamente m funciones, tales que, $\Theta_i : E \to R$ donde m es el número de clases. Es decir, cada función a aprender retorna para cada *ejemplo* m un valor real pi. Cada uno de estos valores pi se denomina probabilidad de la clase i y significa el grado de certeza de que un *ejemplo* sea de la clase i. Idealmente, si además se cumple que $\forall pi : 0 \leq pi \leq 1$, y además, $\sum pi = 1$, estas pi representan *la probabilidad* de que un *ejemplo* sea de la clase i. El conjunto de funciones aprendidas se denomina *estimador de probabilidad.*

Ejemplo. Teniendo el problema de clasificar entre varios medicamentos ¿cuál es el mejor para una determinada patología?. Proporcionar la probabilidad de que sea cada uno de los medicamentos.

b) Categorización. Aquí no se trata de *aprender* una función, sino una *correspondencia.* Es decir, cada *ejemplo* de $\delta = \{\langle e,s\rangle : e \in E, s \in S\}$ así como la correspondencia a aprender $\lambda : E \to S$, pueden asignar *varias* categorías a un mismo e, a diferencia de *la clasificación*, que sólo asigna una y sólo una.

La categorización se puede presentar también en forma de categorización suave (cada categoría asignada va acompañada de su certeza) o en forma de un estimador de probabilidades (se estima una probabilidad para todas las categorías), en este caso la suma de probabilidades puede ser mayor que uno. Por ejemplo, en un conjunto de perfiles de clientes, determinar los productos que puedan comprar.

c) Preferencias o priorización. Este aprendizaje consiste en determinar a partir de dos o más *ejemplos*, un orden de preferencia. Cada *ejemplo* es en realidad una secuencia: $\langle e1, e2, ..., ek \rangle, ei \in E, k \geq 2$, donde el orden de la secuencia representa la predicción. Un conjunto de datos para este problema es, un conjunto de secuencias $\delta : \{\langle e1, e2, ..., ek \rangle : ei \in E\}.$

Ejemplo. De una serie de candidatos para un trabajo, dar un orden priorizado para cubrir el puesto (el modelo de preferencia se habrá estimado a partir de selecciones anteriores (priorizaciones) o comparaciones de grupos de candidatos anteriores, etc).

Quizá lo más característico de esta tarea es la presentación de los datos, ya que, por ejemplo, con *un clasificador suave o un estimador de probabilidad* también se pueden hacer priorizaciones, aunque aquí lo que se prioriza es la clase no el *ejemplo* completo.

d) **La Regresión** *(interpolación o estimación)*. Es una tarea predictiva que consiste en aprender una función que asigna a cada instancia un valor real. Esta es la principal diferencia respecto a la clasificación; el valor a predecir es numérico.

El conjunto de evidencias son correspondencias entre dos conjuntos, $\delta : E \to S$, donde S es el conjunto de valores de salida. El objetivo es aprender una función $\lambda : E \to S$ que represente la correspondencia existente en los *ejemplos*, es decir, para cada valor de E tenemos un único valor para S. La principal diferencia respecto a *la clasificación* es que S es numérico, es decir, puede ser un valor entero o real.

Ejemplos. La estimación de las ventas para el año 2007, predecir el número de unidades defectuosas de una partida de productos, etc.

Tareas descriptivas

Su objetivo es la descripción de los datos existentes. Dentro de estas tareas se encuentran: *el agrupamiento (clustering), las reglas de asociación y el análisis correlacional.* Aquí los *ejemplos* se presentan como un conjunto $\delta = \{e : e \in E\}$, sin etiquetar ni ordenar de ninguna manera.

a) Agrupamiento o *clustering*. También conocido como *segmentación o aglomeración, e*s la tarea descriptiva por excelencia y consiste en obtener grupos a partir de los datos. Aquí se habla de grupos y no de clases, porque a diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo. Al agrupamiento suele llamarse segmentación, ya que parte o segmenta los datos en grupos que pueden ser o no disjuntos. El agrupamiento está muy relacionado con la sumarización, en la que cada grupo formado se considera como un resumen de los elementos que lo forman para así describir de una manera concisa los datos.

La función a obtener es idéntica a la de *la clasificación* $\lambda : E \to S$ con la diferencia de que los valores de S *y* sus miembros se crean o inventan, durante el proceso de aprendizaje.

Ejemplo. Una librería que ofrece sus servicios a través de la red usa el agrupamiento para identificar grupos de clientes con base a sus preferencias de compras que le permita proporcionar un servicio personalizado. Así, cada vez que un cliente se interesa por un libro, el sistema identifica a qué grupo pertenece y le recomienda otros libros comprados por clientes de su mismo grupo.

b) Correlaciones y factorizaciones. Son una tarea descriptiva que su objetivo es examinar el grado de similitud de los valores de dos variables numéricas, es decir, determinar la relevancia de atributos, detectar atributos redundantes o dependencias entre éstos, o seleccionar un subconjunto. Dados los *ejemplos* del conjunto $E = A1xA2x...xAn$, su objetivo es ver si dos o más atributos numéricos Ai y Aj están correlacionados linealmente o de algún otro modo.

Ejemplo. Un inspector de incendios que desea obtener información útil para la prevención de incendios probablemente esté interesado en conocer correlaciones entre el empleo de distintos grosores de protección del material eléctrico y la frecuencia de ocurrencia de incendios.

c) Reglas de asociación (*link analysis*). Su objetivo es similar al anterior, pero para atributos nominales o categóricos. Las reglas de asociación no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados.

Dados los ***ejemplos*** del conjunto $E = A1xA2x...xAn$, una regla de asociación se define generalmente de la siguiente forma:

"si $Ai = a \wedge Aj = b \wedge ... \wedge Ak$ entonces $Ar = u \wedge As = v \wedge ... \wedge Az = w$*"*, donde todos los atributos son nominales y las igualdades se definen utilizando algún valor de los posibles para cada atributo.

Esta tarea se utiliza frecuentemente en el análisis de los hábitos de compra en supermercados, para identificar productos que son comprados de manera simultánea, (el 98% de la gente que compra pañales también compra comida para bebé); esta información se utiliza para inventarios, organización física del almacén o en campañas publicitarias. Existen muchas variantes de las reglas de asociación: las negativas, las secuenciales, multinivel, etc.

d) Dependencias funcionales. Éstas consideran todos los posibles valores (a diferencia del anterior). Se definen de la siguiente manera: "dados los valores de $Ai, Aj, ..., Ak$ puedo determinar el valor de Ar". **Ejemplo.** "Teniendo la edad, el nivel de ingresos, el código postal, el estado civil, se puede determinar con bastante fiabilidad si el cliente tiene vehículo".

e) Detección de valores e instancias anómalas (*outlier detection*). Este método es muy útil para descubrir comportamientos anómalos, que pueden sugerir fraudes, intrusos o comportamientos diferenciados. Su objetivo es encontrar aquellas instancias que no son similares a ninguna (o muy pocas) de las otras instancias.

Generalmente se agrupan los *ejemplos* y se observan aquellas instancias que quedan "desplazadas" de los grupos mayoritarios.

Para ello, se usan *los agrupadores suaves o los estimadores de probabilidad de agrupamiento* con todos los grupos se puede considerar un caso "aislado" y, por tanto, anómalo. **Ejemplo.** Encontrar de las compras realizadas con tarjetas, aquellas que sean anómalas, detección de lavado de dinero, etc.

2.3.2 Tipos de modelos en minería de datos

Los modelos son una descripción de los patrones y relaciones entre los datos, pueden usarse para hacer predicciones, para entender mejor los datos o para explicar situaciones pasadas.

Cada tipo de tarea de minería de datos genera un tipo de modelo específico, los cuales pueden ser:

a) ***Predictivos.*** Estiman valores futuros o desconocidos de variables de interés (variables objetivo o dependientes), utilizando otras variables referidas como independientes o predictivas.

b) ***Descriptivos.*** Identifican patrones que explican o resumen los datos a través de explorar las propiedades de los datos examinados, no para predecir datos nuevos.

2.3.3 Métodos

Cada una de las técnicas o métodos resuelven tipos de tareas específicos y generan un modelo propio, dentro del conjunto de técnicas se encuentran: las técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, programación lógica inductiva y varios tipos de métodos basados en núcleos, entre otros. Cada una de estas técnicas incluye diferentes algoritmos y variaciones de los mismos, así como otro tipo de restricciones que hacen que la efectividad del algoritmo dependa del dominio de aplicación.

2.3.4 Algoritmos en la minería de datos

Los algoritmos de minería de datos se clasifican en dos grandes categorías:

a) Supervisados o predictivos y

b) No supervisados o de descubrimiento del conocimiento.

Los *algoritmos supervisados o predictivos.* Predicen el valor de un atributo (etiqueta) de un conjunto de datos conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otras series de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Los métodos usados en este análisis son los *de clasificación y predicción,* la eficiencia de estos métodos se evalúa de acuerdo a los siguientes criterios:

- **Eficiencia en la predicción.-** Es la habilidad del modelo para predecir correctamente la clase del nuevo conjunto de datos.
- **Rapidez.-** Se trata del costo computacional involucrado en generar y usar el modelo.
- **Robustez.-** Es la habilidad del modelo para hacer predicciones correctas en datos con ruido o valores erróneos o faltantes.
- **Escalabilidad.-** Corresponde a la habilidad de construir el modelo eficientemente dado un gran volumen de datos.
- **Interoperabilidad.-** Se refiere al nivel de comprensión que provee el modelo.

Algoritmos No supervisados o de descubrimiento del conocimiento*.* Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos no supervisados o de descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas.

La Tabla 2.1 muestra algunas de las técnicas de minería de ambas categorías.

Tabla 2.1 Clasificación de algoritmos en minería de datos

Supervisado	No supervisado
▪ Árboles de decisión ▪ Introducción neuronal ▪ Regresión ▪ Series temporales	▪ Detección de desviaciones ▪ Segmentación ▪ Agrupamiento (clustering) ▪ Reglas de asociación ▪ Patrones secuenciales

2.3.5 Relación entre las tareas, métodos y algoritmos

La Tabla 2.2 y la figura 2.6 al final del capítulo, presentan la relación existente entre *las tareas y los métodos* de minería de datos, así como sus *algoritmos* correspondientes para la generación de los modelos.

Tabla 2.2 Descripción de las tareas, métodos y algoritmos

Método o Técnica	Descripción	Algoritmos Asociados	TAREAS				
			Predictivas*		Descriptivas*		
			C	R	A	R /A	C/F
1.- Árboles de decisión y Sistema de aprendizaje de reglas.	Son técnicas fáciles de usar, utilizan atributos discretos y continuos, manejan bien los atributos no significativos, los faltantes y el ruido. Se basan en dos tipos de algoritmos: 1)"Divide y vencerás", como el CART o el ID3/C4.5. 2) "Separa y vencerás", como el CN2.	Árboles de decisión ID3, C4.5, C5.0	✔				
		Árboles de decisión CART	✔	✔			
		Otros árboles de decisión	✔	✔	✔	✔	
		CN2 rules (cobertura)	✔			✔	
2.- Redes neuronales artificiales.	Aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas. La topología de la red y los pesos de las conexiones determinan el patrón aprendido. Encontramos *el perceptrón simple, redes multicapa, redes de base radial, Kohonen, etc.*	Redes Neuronales	✔	✔	✔		
		Redes de *Kohonen*			✔		
3.- Técnicas algebraicas y estadísticas o paramétricas.	Expresan modelos y patrones mediante fórmulas algebraicas, funciones lineales y no lineales, medias, varianzas, correlaciones, etc. Estas técnicas cuando obtienen un patrón, lo hacen a partir de un modelo ya predeterminado del cual, se estiman coeficientes o parámetros.	Regresión Lineal y Logarítmica		✔			✔
		Regresión Logística	✔			✔	

* C=Clasificación, R=Regresión, A=Agrupamiento, R/A=Reglas de Asociación, C/F=Correlaciones y Factorizaciones.

Técnica o Método	Descripción	Algoritmos Asociados	TAREAS				
			Predictivas*		Descriptivas*		
			C	R	A	R/A	C/F
4.- Conteos de frecuencias y tablas de contingencias.	Cuentan la frecuencia en la que dos o más sucesos se presenten conjuntamente.	A priori y similares				✔	
5.- Técnicas Bayesianas	Estiman la probabilidad de pertenencia, mediante la estimación de las probabilidades condicionales inversas o a priori, utilizando el Teorema de *Bayes*.	El clasificador bayesiano *Naive*, y los métodos basados en Máxima Verosimilitud y el algoritmo EM (*Expectation Maximization*).	✔				
6.- Técnicas basadas en casos, en densidad o distancia.	Se basan en distancias al resto de elementos, ya sea directamente, como los vecinos más próximos, de una manera más sofisticada, mediante la estimación de funciones de densidad.	Vecinos más próximos	✔	✔	✔		
		Twostep, Cobweb			✔		
		Kmeans			✔		
7.- Análisis Factorial y de componentes Principales o Método de *Karhunen-Loeve.*	El análisis de componentes principales, reduce la dimensionalidad por transformación de los atributos o variables originales $x1,x2,...,xm$ de los *ejemplos* en otro conjunto de atributos $f1,f2,...,fp$, donde $p \leq m$.	Mínimos Cuadrados, Máxima Verosimilitud, Factorización de ejes principales, entre otros.					✔
8.- Técnicas estocásticas y difusas.	Aquí se incluyen la mayoría de las técnicas que, junto a las redes neuronales, forman la computación flexible (*soft computing*). Los componentes aleatorios son fundamentales como el *simulated annealing*, los métodos evolutivos y genéticos, o las funciones de pertenencia difusas.	Algoritmos genéticos y evolutivos	✔	✔	✔	✔	✔
9.- Núcleo y máquinas de soporte vectorial	Intentan maximizar el margen entre los grupos o las clases formadas. Para ello se basan en transformaciones, llamadas núcleos (*kernels*) que pueden aumentar la dimensionalidad.	Máquinas de vectores de soporte	✔	✔	✔		

* C=Clasificación, R=Regresión, A=Agrupamiento, R/A=Reglas de Asociación, C/F=Correlaciones y Factorizaciones.

Un aspecto muy relevante en la extracción de conocimiento es que los modelos extraídos sean comprensibles. Es importante señalar que algunas de las técnicas descritas anteriormente resuelven tareas *sin construir* (explícitamente) modelos.

Los métodos sin modelo y con modelo reciben generalmente el nombre de *métodos retardados o perezosos (lazy) y métodos anticipativos o impacientes (eager).*

Métodos retardados o perezosos. El método actúa para cada pregunta o predicción requerida. No se construye modelo. Optimización local. Los *ejemplos* deben preservarse porque son necesarios para realizar cada predicción. El tiempo de respuesta empieza a degradarse cuando el número de *ejemplos* es muy grande porque hay que consultar muchos de ellos. La ventaja, es que no hay que entrenar al modelo.

Métodos anticipativos o impacientes. El método obtiene un modelo a partir de todos los *ejemplos.* Los *ejemplos*, por tanto, pueden ignorarse. Optimización global. Se requiere un tiempo de entrenamiento, que suele ser grande, pero una vez entrenado el modelo, su aplicación suele ser instantánea.

La Tabla 2.3 presenta la clasificación de las técnicas que se encuentran dentro de los métodos retardados o perezosos y de los métodos anticipativos o impacientes.

Tabla 2.3 Métodos anticipativos y retardados

Tipo	**Con modelo** (Útiles para extracción de conocimiento)	**Sin modelo o no inteligible**
Anticipativo	▪ Regresión lineal ▪ K-medias ▪ ID3, C4.5, **CART** ▪ CN2 ▪ A priori ▪ ILP, IFLP ▪ Redes bayesianas ▪ Reglas difusas	▪ Redes neuronales ▪ Radial basis functions ▪ Clasificador bayesiano naïve ▪ Máquinas de vectores de soporte ▪ Boosting
Retardado		▪ k-NN (Vecinos más próximos) ▪ CBR (Case-based reasoning)

No todos los métodos con modelo son comprensibles, basta examinar a una red neuronal entrenada e intentar extraer algún conocimiento sobre los patrones encontrados. En cambio, un árbol de decisión con pocas reglas puede ser mucho más sencillo de examinar, de comprender y, por tanto, de validar, modificar y adaptar a las necesidades de aplicación.

2.4 Software para minería de datos

Los algoritmos anteriores, se encuentran en software especializado, el cual de acuerdo a su funcionalidad, se clasifica en: *librerías, suites y herramientas específicas.*

a) Librerías.- Son un conjunto de métodos que implementan las utilidades básicas de la minería de datos, como pueden ser: acceso a datos, inferencia de modelos *(árboles de decisión, redes neuronales, métodos bayesianos, etc)*, exportación y comprobación de resultados, etc.

b) Suites.- Estas integran en un mismo entorno capacidades para el preprocesado de datos, diferentes modelos de análisis, facilidades para el diseño de experimentos y soporte gráfico para visualizar los resultados. Tienen una interfaz gráfica que facilita la interacción entre el usuario y la herramienta. La mayoría cuentan con arquitectura cliente-servidor y algunas son de libre distribución. Algunas de las más representativas son: *SPSS Clementine, WEKA, KEPLER, ODMS (Oracle Data Mining Suite- Darwin), DBMiner, YALE, DB2 Intelligent Miner (IBM), SAS Enterprise Miner, Statistica Data Miner,* entre otras.

c) Herramientas específicas.- Se caracterizan por centrarse en un determinado modelo (redes neuronales, árboles de decisión, modelos estadísticos, etc) o en una determinada tarea de minería de datos (clasificación, agrupamientos, etc). No requieren conocimientos previos de programación. Ejemplos: CART, AutoClass, Neural Planner, NeuroDiet y Easy NN-Plus, NeuroShell, See5/C5.0.

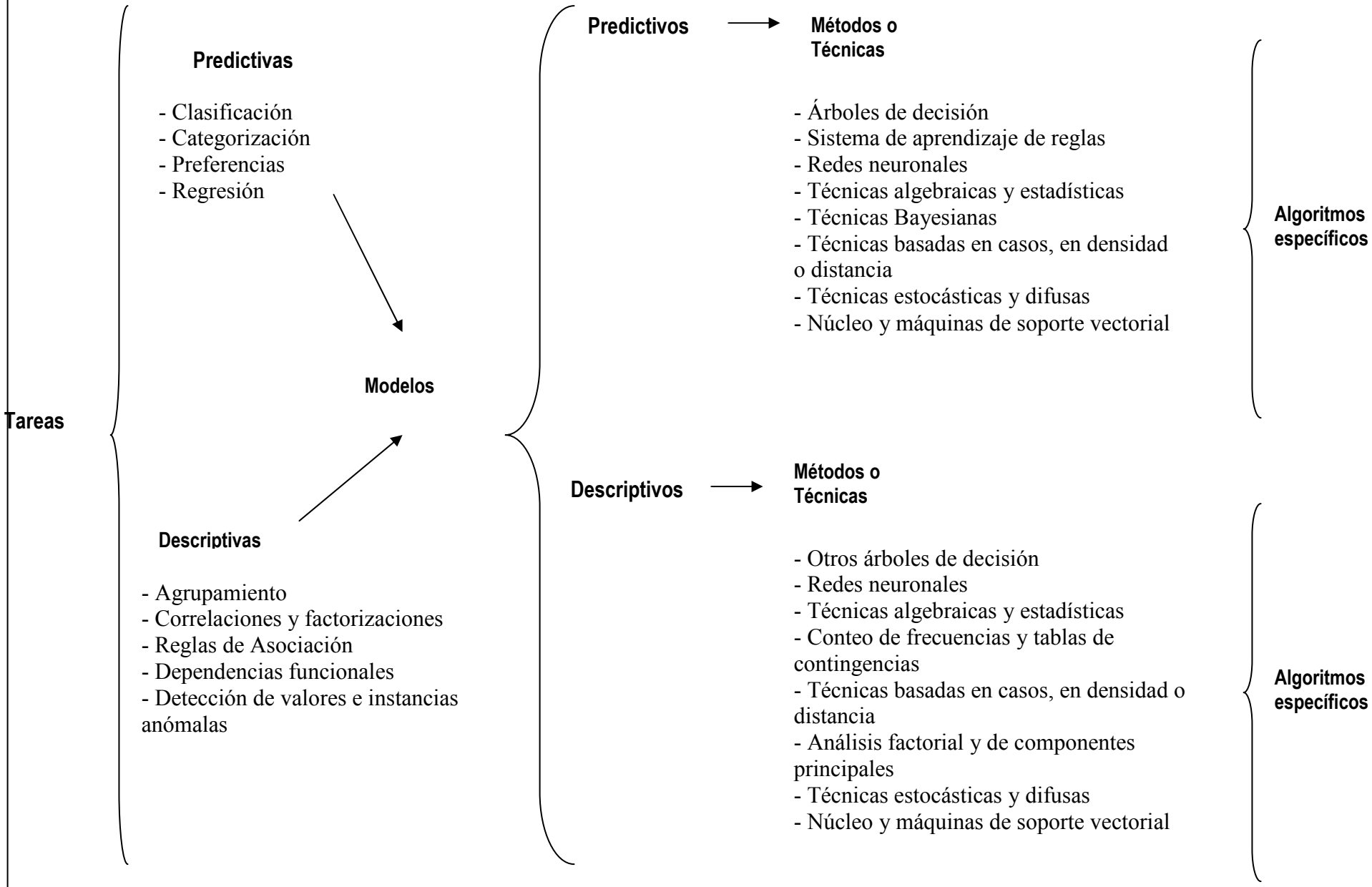


Figura 2.6 Esquema general de la relación entre tareas, modelos, métodos y algoritmos de la minería de datos

Capítulo 3

ÁRBOLES DE DECISIÓN

Palabras Clave: Árbol de decisión. Árbol de clasificación. Algoritmo CART.

3.1 Antecedentes

El uso de árboles de decisión tuvo su origen en las ciencias sociales con los trabajos de Sonquist y Morgan en 1964, Morgan y Messenger en 1979 realizados en el Survey Research Center del Institute for Social Research de la Universidad de Michigan. En el área Estadística, Breiman, Friedman, Olshen y Stone durante 1984 introdujeron nuevos algoritmos para construcción de árboles y los aplicaron a problemas de regresión y clasificación. Casi al mismo tiempo el proceso de inducción mediante árboles de decisión comenzó a ser usado por la comunidad de Aprendizaje automático (Machine Learning) representado por Michalski en 1973, Quinlan en 1983 y la comunidad de Reconocimiento de patrones (Pattern Recognition) por Henrichon y Fu, en 1969. Hoy en día la comunidad que más contribuye al desarrollo de clasificación basada en árboles de decisión es la de Aprendizaje automático.

3.2 Definición

Un *árbol de decisión* se define como "un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplan desde la raíz del árbol hasta alguna de sus hojas" [20]*. Ver figura 3.1.*

De manera general, se dice que la representación gráfica de un árbol es un diagrama con estructura parecida a la de un árbol invertido, donde cada nodo del árbol representa un atributo de los registros del problema, debiendo partir de un nodo raíz (el nodo de hasta arriba en el árbol), donde dependiendo del algoritmo utilizado, se realiza la mejor discriminación sobre el punto de partida de clasificación o predicción que se desea obtener, es decir, se realiza una prueba (pregunta sobre un atributo concreto) al registro en análisis con respecto al nodo raíz, después se continúa con un *nodo hijo*, el cual a su vez también realiza una prueba al registro y así sucesivamente hasta llegar a un nodo final llamado *nodo hoja*.

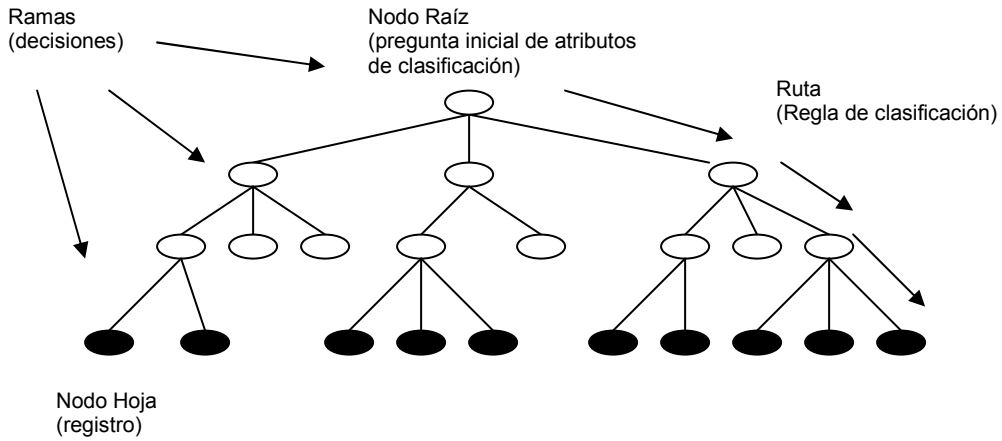


Figura 3.1 *Árbol de decisión mezclado, con ramificaciones binarias y ternarias*

[20] Hernández Orallo, J., Ramírez Quintana, M.J. & Ferri Ramírez. (2004). *Introducción a la Minería de Datos.* Madrid, España: Pearson Educación. 680 pp.

Más formalmente, la técnica de árbol de decisión:

1.- Reúne una gran cantidad de *ejemplos* (datos).

2.- Forma dos conjuntos disjuntos: entrenamiento y prueba.

3.- Usa el algoritmo de aprendizaje para generar una hipótesis *H.*

4.- Mide el porcentaje de clasificación correcta de *H* en el conjunto de prueba.

5.- Repite los pasos del 1 al 4 para diferentes tamaños de conjuntos de entrenamiento y diferentes conjuntos seleccionados aleatoriamente.

Realizando las pruebas correctas a los conjuntos se pueden encontrar clasificaciones muy interesantes en muy pocos niveles del árbol.

Como forma de representación del conocimiento, los árboles de decisión destacan por su sencillez y su dominio de aplicación no está restringido a un ámbito concreto sino que pueden utilizarse en diversas áreas como el diagnóstico médico, los juegos, la predicción meteorológica, el control de calidad, los procedimientos legales, etc.

Ejemplo. En un hospital, cuando se admite un paciente con “ataque al corazón”, se miden 19 variables durante las primeras 24 horas, algunas variables como la presión sanguínea, la edad y otras 17 variables binarias resumen los síntomas médicos considerados como indicadores importantes de la condición del paciente. El objetivo de un estudio reciente, fue *el desarrollar una metodología para identificar a los pacientes de alto riesgo*, es decir, aquellos que no sobrevivirán menos de 30 días sobre la base de los datos obtenidos durante las primeras 24 horas.

En la figura 3.2, se muestra *el árbol de clasificación* que se produjo en el estudio.

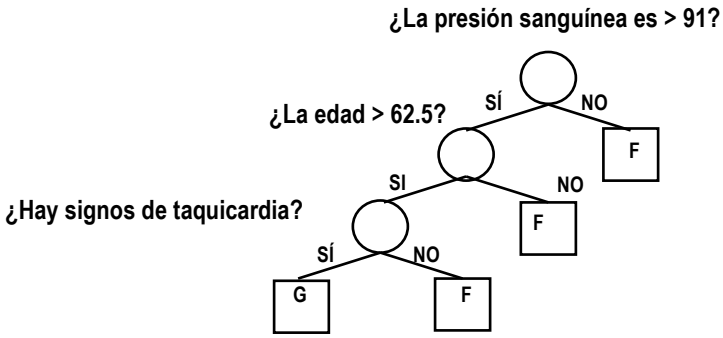


Figura 3.2 *Árbol de decisión, para pacientes con "ataque al corazón"*

Este árbol de decisión funciona como un *clasificador,* es decir, dado un nuevo individuo lo clasifica en una de las dos clases posibles: "F = no tiene alto riesgo" o "G = alto riesgo".

Los árboles de decisión se utilizan principalmente para tareas de *clasificación y de regresión,* aunque su difusión ha hecho que también se utilicen en las tareas de *agrupamiento, y estimación de probabilidades,* así mismo existen *hibridaciones* con otros métodos para explotar la técnica, lo que hace que exista una gran variedad de algoritmos que utilizan una o varias tareas, haciendo las adecuaciones pertinentes para cada caso. Las principales diferencias entre los algoritmos de construcción de árboles de decisión radican *en las estrategias de poda y en la regla adoptada para particionar nodos.*

Los árboles de decisión trabajan con dos tipos de variables: *las variables de criterio o dependientes y las variables predictoras o independientes.* Las variables criterio son aquellas cuyos resultados se desean predecir a partir de otras variables. Las variables predictoras son las que predicen el patrón de la variable criterio.

3.3 Metodología de los árboles de clasificación[21]

El árbol de clasificación, es uno de los métodos de aprendizaje inductivo supervisado más utilizados. Dependiendo del problema, el propósito básico de un estudio de clasificación será, no sólo el de producir "clasificadores precisos y veraces", sino también el de *proveer un conocimiento profundo acerca de la estructura predictiva de los datos.*

Entre los algoritmos clasificadores basados en árboles, el que se escogió para esta investigación es el CART versión 5.0 (acrónimo de *Classification And Regression Trees*).

El CART versión 5.0, es un software creado por los laboratorios Salford Systems, y es el único que se basa en el código original CART, desarrollado por Brieman, y otros de las universidades de Stanford y Berkeley en California.

3.3.1 Estructura de árboles binarios

La estructura de los *árboles binarios* se construye realizando divisiones repetidas de subconjuntos de un vector de *ejemplos* X, formando dos subconjuntos descendentes, empezando con el X mismo.

[21] Algunos de los conceptos y terminología utilizada en la definición y construcción de los árboles de clasificación se expuso en el segundo capítulo.

Ejemplo. Árbol de 6 clases.

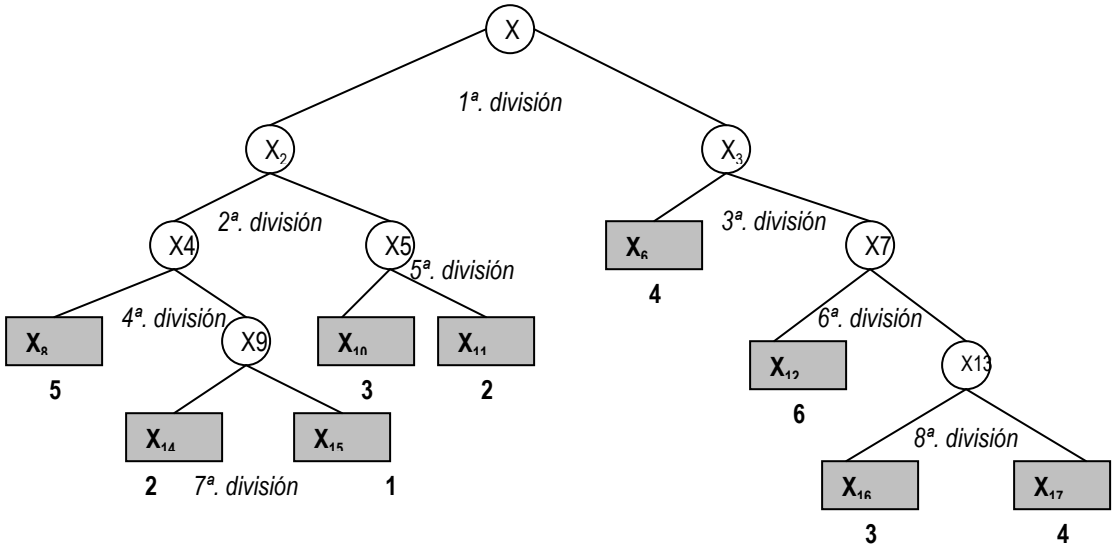


Figura 3.3 *Ejemplo de árbol binario*

En la figura 3.3, se observa que X_2 y X_3 son disjuntos, con $X = X_2 \cup X_3$. De forma similar, X_4 y X_5 son disjuntos con $X_2 = X_4 \cup X_5$, y $X_3 = X_6 \cup X_7$. Aquellos subconjuntos que no son divididos, como son X_6, X_8, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{16} y X_{17}, son llamados *subconjuntos de nodos terminales o nodos hoja*, indicados con un rectángulo, y los *subconjuntos de nodos no terminales o nodos intermedios* están representados por círculos.

Cada *nodo terminal* está designado por una *etiqueta de clase*, puede haber dos o más *nodos terminales* con una misma *etiqueta de clase.*

La *partición* correspondiente al *clasificador,* se obtiene juntando todos los *nodos terminales* correspondientes a la misma *clase*, de esta forma, se tiene:

Clase 1 = X_{15} Clase 2 = $X_{11} \cup X_{14}$

Clase 3 = $X_{10} \cup X_{16}$ Clase 4 = $X_6 \cup X_{17}$

Clase 5 = X_8 Clase 6 = X_{12}

Las divisiones o ramificaciones se forman por las condiciones en las coordenadas de $X = (x_1, x_2, ...)$. Por ejemplo. La primera ramificación de X en X_2 y X_3 puede ser: $X_2 = \{x; x_4 \leq 7\}$, $X_3 = \{x; x_4 > 7\}$.

Una vez formadas las ramificaciones, el *árbol clasificador* predice una *clase* para cada una de las medidas del vector X, y cuando X, finalmente se mueve a un *nodo terminal, la clase predecida* está dada por la etiqueta de clase de éste.

El CART y sus variaciones son métodos "divide y vencerás" y se caracteriza fundamentalmente, por *realizar particiones binarias utilizando el criterio denominado GINI y por utilizar una estrategia de poda basada en el criterio de coste-complejidad*, conceptos que se explicarán más adelante en este capítulo.

La metodología CART se resume en:

1.- Aprendizaje. Consiste en la construcción del árbol a partir de un conjunto de *ejemplos.* Constituye la fase más compleja y la que determina el resultado final.

2.- Clasificación. Consiste en el etiquetado de un patrón, X, independiente del conjunto de aprendizaje. Se trata de responder a las preguntas asociadas a los nodos interiores utilizando los valores de los atributos del patrón X. Este proceso se repite desde el nodo raíz hasta alcanzar una hoja, siguiendo el camino impuesto por el resultado de cada evaluación.

Antes de empezar con la construcción del árbol paso por paso, es necesario precisar algunos conceptos.

Para definir a *un clasificador o una regla de clasificación,* se debe contar con un vector de medidas X, como en el ejemplo de *los pacientes con ataques al corazón*, el conjunto de 19 medidas involucradas en el caso, se organizan como: $x_1, x_2, ...$ donde, x_1, *es la edad,* x_2, *es la presión sanguínea,* etc. Una vez teniendo, las medidas $(x_1, x_2, ...)$, se define *el vector de medidas* X como un espacio de 19 dimensiones, tal que la primera coordenada *(edad),* toma valores enteros de 0 a 200, la segunda coordenada, *la presión sanguínea,* puede ser definida en un rango continuo de 50 a 150, etc.

Supongamos que los casos o *ejemplos* caen en J clases. El número de clases está dado por $1,2,..., j$ y C será el conjunto de clases, esto es $C = \{1,..., J\}$.

Una forma de predecir los miembros de una clase es, construyendo *una regla* que asigne un miembro de clase contenido en C para cada medida del vector x en X. Esto es, dado cualquier $x \in X$, la regla asigna una de las clases $\{1,..., J\}$ a x.

La estructura del vector X, puede contener variables tanto numéricas como categóricas.

De lo anterior, se deduce:

Definición:"Un clasificador o regla de clasificación, es una función $d(x)$ definida sobre X, tal que para cada x, $d(x)$ es igual a uno de los números $1,2,...,J$."[22]

Los clasificadores, se construyen sobre la base de la experiencia de los hechos ocurridos, lo que se conoce como *muestra de aprendizaje.*

Definición*:* "Una muestra de aprendizaje consiste en datos $(x_1, j_1),...,(x_N, j_N)$ en N casos, donde $x_n \in X$ y $j_n \in \{1,...,J\}, n = 1,...,N.$ La muestra de aprendizaje se denota por L; es decir, $L = \{(x_1, j_1),...,(x_N, J_N)\}$."[23]

3.3.2 Construcción del árbol de clasificación

La construcción del árbol presenta el siguiente esquema recursivo:

1. El avance está basado en la partición de un nodo de acuerdo a alguna regla, normalmente evaluando una condición sobre el valor de alguna variable:

   Los ejemplos que verifican la condición se asignan a uno de los dos nodos hijo y los restantes al otro.

Al separar los *ejemplos* en distintos hijos, la cardinalidad de los nodos irá disminuyendo a medida que se desciende en el árbol, entendiendo por cardinalidad, el número total de *ejemplos de entrenamiento* que caen en ese nodo.

[22] Breiman, Leo. & Friedman, Jerome H. (1984). *Classification and Regression Trees.* Chapman & Hall.

[23] *Idem.*

2. La condición de parada tiene como objetivo detener el proceso de partición de nodos. En ocasiones, se poda el árbol resultante utilizando alguna regla de poda.

La longitud de un árbol de decisión estará determinada por el número de subconjuntos que se requieran formar a través de los nodos anteriores a los nodos hoja, es decir, se tendrán tantos niveles como se requieran hasta obtener una *clasificación* adecuada según el problema que se esté tratando.

Ejemplo. En este ejemplo[24] se muestra el proceso para construir un árbol a partir de un conjunto de *ejemplos* pertenecientes a tres clases y con 25 atributos con valores no categóricos. El conjunto de aprendizaje consta de $N = 300$ *ejemplos* de manera que $Ni = 100,\ i = 1,2,3$.

1. Construcción del nodo raíz

Inicialmente se asignan todos los ejemplos a la raíz (ver figura 3.4) de manera que éste contiene 100 ejemplos de cada clase. Esta situación, en la que todas las clases están igualmente representadas, corresponde a la situación de máxima impureza, es decir, ninguna clase "domina" sobre las otras.

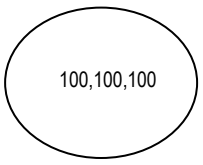


Figura 3.4 *Nodo raíz del árbol*

[24] Recuperado el 19 de julio del 2006, de *http://www-etsi2.ugr.es/depar/ccia/rf/www/tema3_00-01_www/node26.html*

2. Particionar el nodo raíz

Se trata de seleccionar la mejor partición del nodo raíz entre todas las posibles. Este proceso puede descomponerse en tres pasos:

2.1 Examinar *todas* las particiones de la forma $¿X_1 < C?$ donde:

$$\min(X_1) \leq C \leq \max(X_1)$$

Por ejemplo, sea C = 1.1. Los *ejemplos* para los que se verifica que $X_1 < 1.1$ van al nodo izquierdo, y los otros al derecho (ver figura 3.5).

Una vez examinadas todas las particiones para la variable X_1, se considera la mejor partición asociada a esta variable. Por ejemplo, sea ésta $¿X_1 < 10.7?$

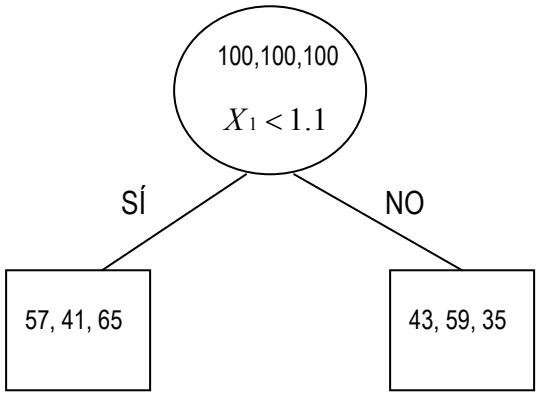


Figura 3.5 *Partición* $¿X_1 < 1.1?$

2.2 Repetir el proceso anterior para $X_2, X_3, ..., X_{25}$.

2.3 Seleccionar la mejor partición entre las mejores de $X_1, X_2, X_3, ..., X_{25}$

Por ejemplo, si la mejor partición se consigue para la variable X_8 y la partición es la asociada a la condición $¿X_8 < 3.2?$, el árbol resultante se muestra en la figura 3.6.

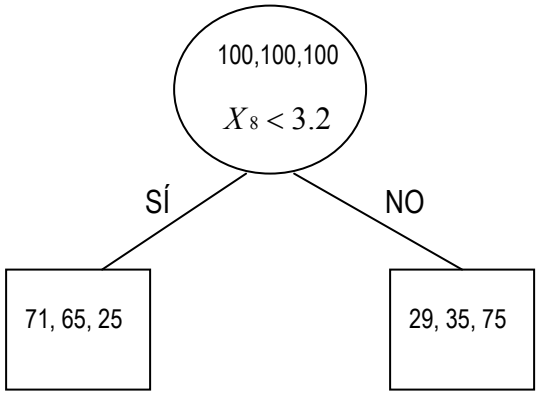


Figura 3.6 *Partición* $¿X_8 < 3.2?$

Si se comparan los árboles de las figuras anteriores, observamos que el primero (figura 3.5), aún siendo más puro que el de la figura 3.4, las proporciones de las clases en cada nodo no son determinantes, en el sentido de que ninguna destaca claramente sobre las otras. En el tercero (figura 3.6) estas proporciones son más determinantes, haciendo que:

a) la clase 3 esté muy por debajo de las clases 1 y 2 en el nodo izquierdo y

b) la clase 3 sea dominante en el nodo derecho.

3. Repetir el paso 2 para los nodos hijo

Por ejemplo, sea $¿X_3 < -0.8?$ la mejor partición para el nodo izquierdo y $¿X_1 < 17.9?$ la mejor para el derecho. En la figura 3.7 se muestra el árbol resultante de estas particiones

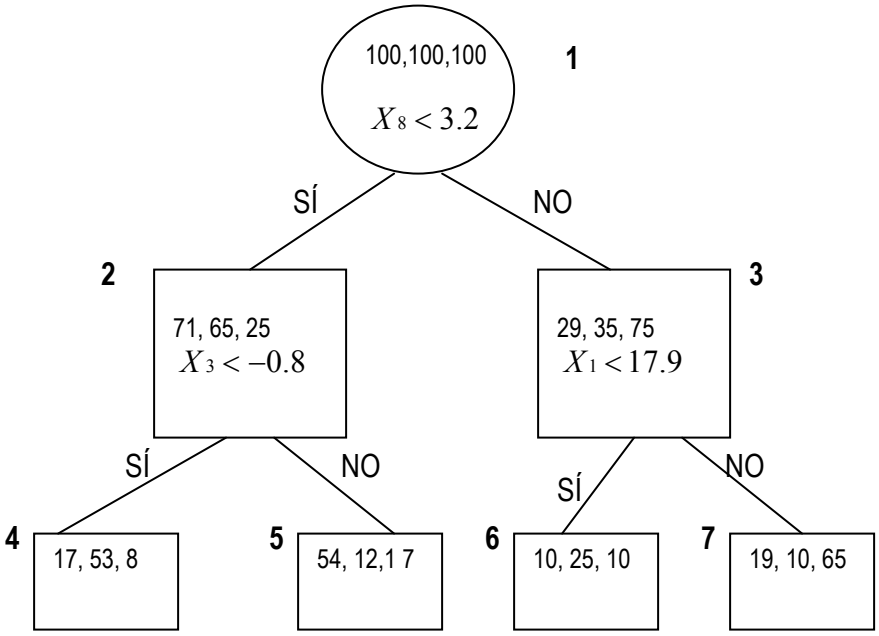


Figura 3.7 *Árbol resultante de partir el árbol de la figura 3.6*

Estas particiones hacen que los nodos 4 y 5 diferencien claramente las clases 2 y 3, respectivamente, mientras que en los nodos 6 y 7 se diferencian las clases 2 y 3, respectivamente. Se observa que las particiones efectuadas han ido "definiendo" una clase mayoritaria en cada nodo resultante, o expresado de otra manera, han ido aumentando la pureza de los nodos asociados a cada partición. Este proceso de división puede continuar para cada uno de los cuatro nodos que se han obtenido o, para cada caso, plantearse si hay que detenerse.

4. Criterio de parada

Establecer el criterio de parada para obtener un buen árbol de decisión no es sencillo. Más adelante se definirá la manera adecuada de hacerlo, basándose en la pureza del nodo.

Un criterio muy simple, puede ser el siguiente: un nodo se declarará terminal, y en consecuencia no se dividirá si la clase dominante tiene más del 60% de los *ejemplos* asociados a ese nodo.

En este ejemplo, y considerando el árbol de la figura 3.7, si $N(t)$ es el número total de *ejemplos* asociados al nodo t y $Ni(t)$ es el número de *ejemplos* de la clase i asociado al nodo t,

- Nodo 4: $N(4) = 78$ 60% de 78=46.8 $N_2(4) = 53$ =>Parar
- Nodo 5: $N(5) = 83$ 60% de 83=49.8 $N_1(5) = 51$ =>Parar
- Nodo 6: $N(6) = 45$ 60% de 45=27 $N_2(6) = 25$ =>Seguir
- Nodo 7: $N(7) = 94$ 60% de 94=56.4 $N_3(7) = 65$ =>Parar

En este caso, se detendría la división de los nodos 4, 5, y 7, mientras que el nodo 6 continuaría su división como se indicó en los pasos 2 y 3. El resultado de este nodo se muestra en la figura 3.8.

Se puede plantear si era necesaria la división de los nodos 1, 2 y 3. Procedemos como para los nodos 4, 5, 6 y 7.

- Nodo 1: $N(1) = 300$ 60% de 300=180. En este nodo no hay clase dominante $(N_i(1) = 100 \quad i = 1,2,3)$ =>Seguir
- Nodo 2: $N(2) = 161$ 60% de 161=96.6 $N_1(2) = 71$ =>Seguir
- Nodo 3: $N(3) = 139$ 60% de 139=83.4 $N_3(3) = 75$ =>Seguir

Así, se hizo bien al dividir estos nodos.

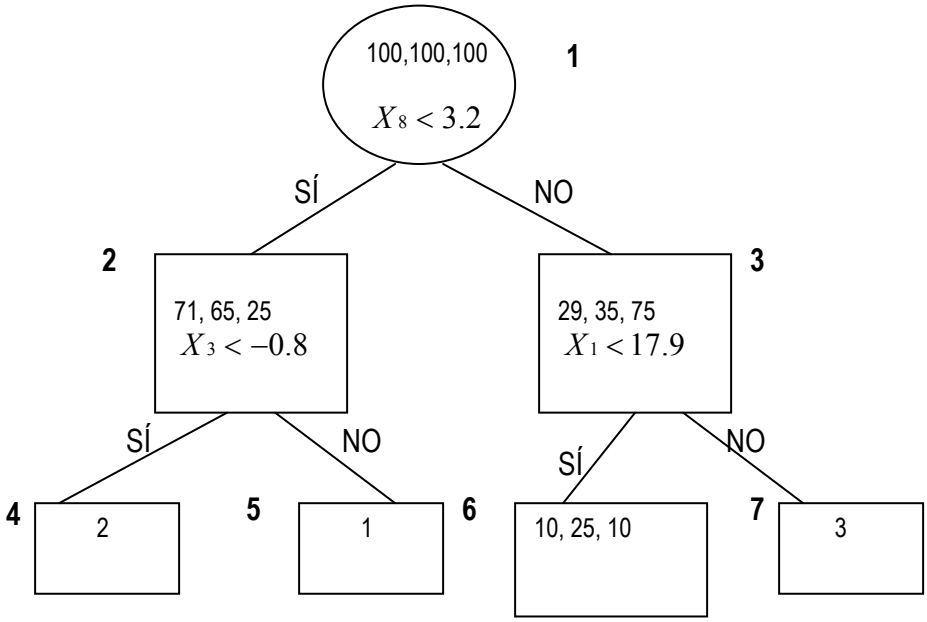


Figura 3.8 *Árbol resultante al declarar los nodos 4,5 y 7 como hojas*

Finalmente, si el resultado de partir el nodo 6 es el mostrado en la figura 3.9, es fácil comprobar que los nodos 6.1 y 6.2 no requieren más particiones (Ver figura 3.10).

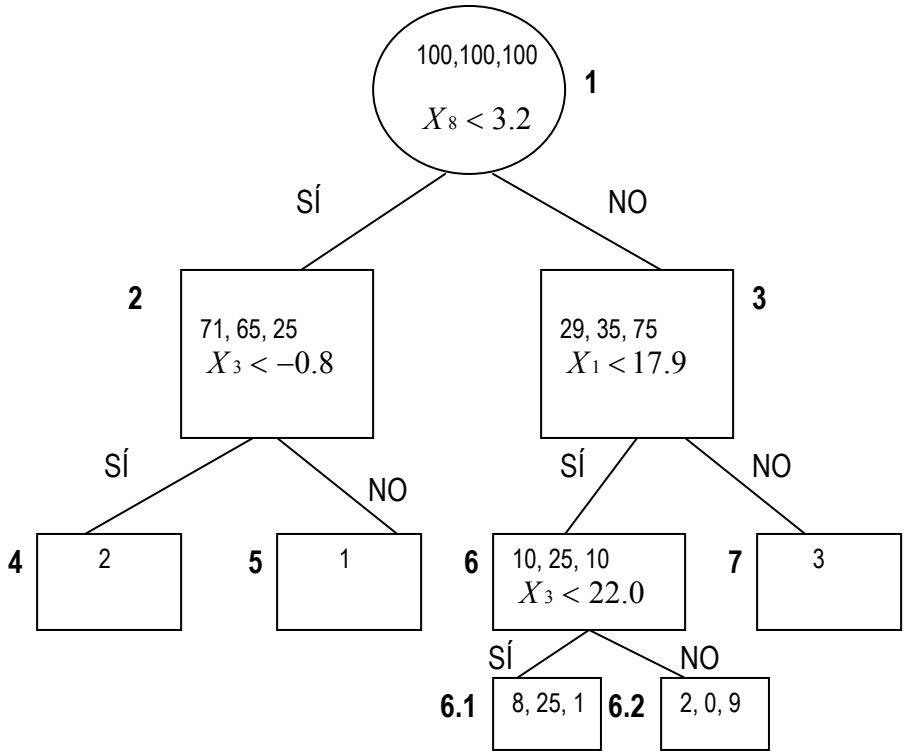


Figura 3.9 *Árbol resultante de partir el nodo 6 del árbol de la figura 3.8*

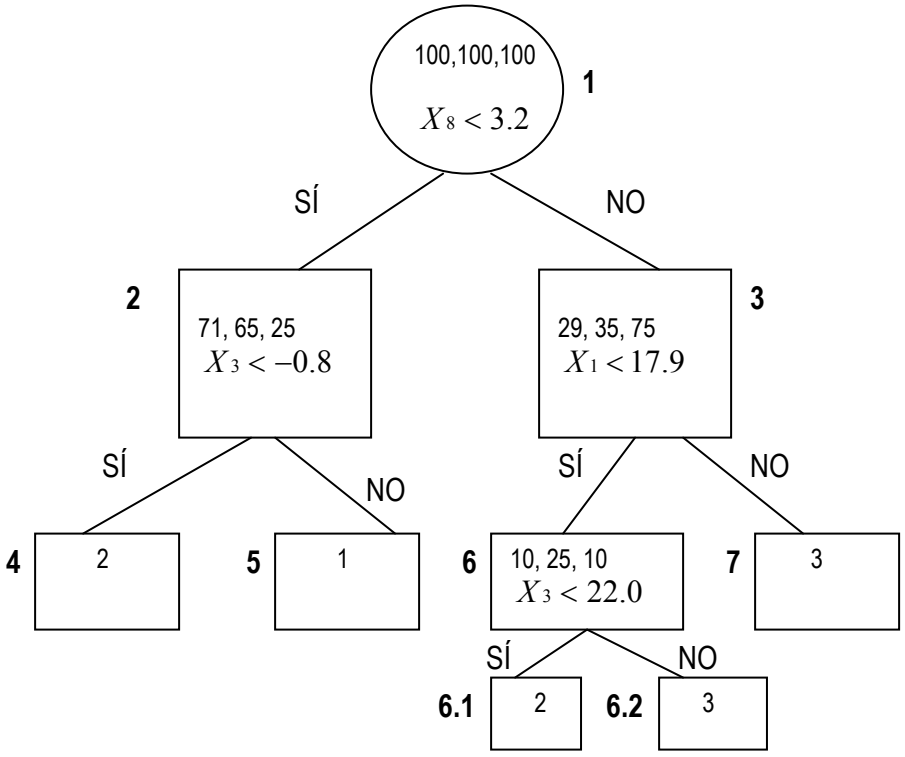


Figura 3.10 *Árbol resultante de partir el nodo 6 del árbol de la figura 3.9*

3.3.3 Selección de las particiones

Dos características muy importantes para que el algoritmo funcione correctamente y que diferencia a los algoritmos de "partición" son: el *número de particiones* a considerar, y el *criterio de selección de particiones.*

a. El número de particiones

Entre más tipos de condiciones se generen, más posibilidades se tienen de encontrar los patrones que están detrás de los datos. Esto da como resultado árboles más expresivos y probablemente más precisos, pero entre más particiones existan, la complejidad del algoritmo aumenta, por lo que debe existir un equilibrio entre expresividad y eficiencia.

Los tipos de particiones que se pueden hallar son *nominales y numéricas.*

b. El criterio de selección de particiones

¿De qué forma se hacen las particiones y se selecciona la mejor de entre las posibles en cada momento?

Una partición divide a un conjunto de *ejemplos* en conjuntos disjuntos. En CART las particiones son binarias, resultado de evaluar una condición que tiene dos únicas respuestas: *Sí o No.*

El objetivo de una partición *es incrementar la homogeneidad (en términos de clase) de los subconjuntos resultantes*, es decir, qué éstos sean más puros que el conjunto originario. Cada partición tiene asociada una medida de pureza, que se utiliza para: la selección de la mejor partición y como criterio de parada.

Formulación de la regla de partición

Se presentará la forma en que CART formula las preguntas y decide sobre cuál es la mejor partición. Primeramente se fijará el marco teórico de trabajo. Sea Q el conjunto de preguntas binarias de la forma:

$$\{¿X \in A?\ \}, A \subset P$$

El conjunto Q genera un conjunto de particiones, s, en cada nodo, t y cada nodo t se particiona en t_L (izquierdo) y t_R (derecho) de manera que:

- Los casos de t que verifican la condición $¿X \in A?$ se asignan a t_L,

$$t_L = t \cap A$$

- Los casos de t que no la verifican se asignan a t_R.

$$t_R = t \cap \overline{A}$$

En CART se define el llamado ***conjunto estándar de preguntas*** de la siguiente forma:

1. Cada partición depende de ***un único*** atributo.

2. Si X_i es un atributo ***categórico***, que toma valores en $\{c_1, c_2, ..., c_L\}$, Q incluye las preguntas:

$$¿X_i \in C?$$

donde C es un conjunto de entre los subconjuntos de $\{c_1, c_2, ..., c_L\}$.

Por ejemplo, si X_2 toma valores en $\{Rojo, Verde, Azul\}$, las preguntas válidas son del tipo $¿X_2 \in \{Rojo\}?,\ ¿X_2 \in \{Verde\}?, etc$.

3. Si X_i es un atributo ***continuo***, Q incluye preguntas del tipo:

$$¿X_i \leq v?$$

donde v es un valor real, teóricamente cualquiera. En CART, no obstante, v es el punto medio de dos valores consecutivos de Xi. Esta heurística[25] simplifica enormemente los cálculos haciendo que los problemas sean tratables.

Por ejemplo, si X_1 es un atributo continuo (real) cuyos valores son: 0.1,0.5 y 1.0 las preguntas válidas que se evalúan son las siguientes:

$$¿X_1 \leq (0.1+0.5)/2?,\ ¿X_1 \leq (0.5+1.0)/2?$$

Como se puede deducir, el conjunto de preguntas que pueden formularse (y que evaluará CART), aunque puede ser muy amplio, está perfectamente definido y es sistemático.

Criterios de partición

Ahora se presentará *¿cómo calcular la mejor partición entre todas las posibles?*

Cada partición tiene asociada *una medida de pureza* y se tratará de incrementar la *homogeneidad* de los subconjuntos resultantes de la partición, esto es, que sean más puros que el conjunto originario.

[25] De acuerdo con ANSI/IEEE Std 100-1984, la heurística trata de aquellos métodos o algoritmos exploratorios para la resolución de problemas en los que las soluciones se descubren por la evaluación del progreso logrado en la búsqueda de un resultado final.

Función de impureza y medida de impureza

Una *función de impureza* es una función Φ definida sobre $J-uplas$ de la forma $(c_1, c_2, ..., c_J)$ tal que:

a) $c_j \geq 0$ para $j = 1,2,...J$ y

b) $\sum_j c_j = 1$, con las siguientes propiedades:

i. Φ tiene un único máximo en $\left(\frac{1}{J}, \frac{1}{J}, ..., \frac{1}{J}\right)$.

ii. Φ alcanza su mínimo en $(1,0,0,...,0),(0,1,0,...,0),...,(0,0,0,...,1)$

y el valor del mínimo es 0.

iii. Φ es una función simétrica de $(c_1, c_2, ..., c_J)$

Relacionada con la función de impureza está *la medida de impureza* de un nodo.

Dada una función de impureza Φ, definimos *la medida de impureza* de cualquier nodo t, y se escribe $i(t)$, como:

$$i(t) = \Phi(p(1|t), p(2|t), ..., p(J|t))$$

donde $p(j|t)$ es la probabilidad de que un caso del nodo t (un ejemplo asociado al nodo t) sea de clase j.

Estas probabilidades pueden calcularse empíricamente como la proporción de casos de clase j en el nodo t:

$$p(j|t) = \frac{N_j(t)}{N(t)}$$

Entonces, *la medida de impureza de un nodo es el resultado de evaluar la función de impureza sobre ese nodo tomando las proporciones relativas de cada clase como los* c_j. Observar que, por un lado,

a) $p(j|t) \geq 0$

b) $\sum_j p(j|t) = \sum_j \frac{N_j(t)}{N(t)} = \frac{1}{N(t)} \sum_j N_j(t) = 1$

lo que garantiza que los componentes de la $J-upla$, calculados en términos de proporción relativa son válidos para evaluar la función de impureza.

Por otro lado:

i. La máxima impureza (mínima pureza) se obtiene cuando todas las clases están igualmente representadas en t.

ii. La mínima impureza (máxima pureza) se obtiene cuando en t sólo hay casos de *una sola clase* (máxima homogeneidad).

iii. Cualquier permutación de los c_j produce el mismo resultado en el valor de impureza. Por ejemplo para dos nodos $t_j \neq t_k,\ i(t_j) = \Phi(0.7,0.2,0.1) = \Phi(0.2,0.1,0.7) = i(t_k)$

Bondad de una partición

La bondad de una partición s en un nodo t debe estar relacionada con la impureza resultantes de la partición, t_L y t_R.

Suponga una partición candidata, s, que divide t en t_L y t_R (ver figura 3.11) de forma que una proporción p_L de los casos de t va a t_L y una proporción p_R va a t_R.

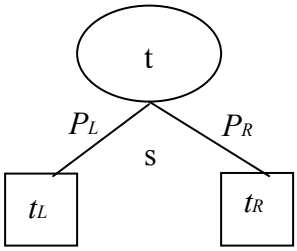


Figura 3.11 *La partición s divide t en t_L y t_R*

La **bondad de la partición s en un nodo t,** $\Phi(s,t)$, se define como *el decrecimiento en impureza* conseguido con ella:

$$\Phi(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Así, como conocemos cómo calcular $i(t)$, podemos calcular $\Phi(s,t)$ para cada partición s y seleccionar la *mejor partición* como la que proporciona la mayor bondad $\Phi(s,t)$.

Para establecer el efecto que produce la selección de la mejor partición en cada nodo sobre el árbol final necesitamos una medida de la impureza global del árbol.

Impureza de un árbol

Supongamos construido un árbol binario T, obtenido mediante una serie de particiones. Sea $\tilde{T}$ el conjunto de nodos terminales del árbol T.

Sea $I(t) = i(t)\,p(t),$ donde $p(t)$ es la probabilidad de que un caso cualquiera esté en el nodo t. La impureza del árbol T, $I(T),$ se define como:

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)\,p(t)$$

En definitiva*, la impureza de un árbol se calcula únicamente con base al conjunto de nodos terminales.*

Una conclusión fundamental de Breiman (1984), es la siguiente:

La selección continuada de las particiones que maximizan $\Delta i(s,t)$ *es equivalente a seleccionar las particiones que minimizan la impureza global* $I(T)$*,*

lo que significa que la estrategia de selección de la mejor partición en cada nodo conduce a la solución óptima considerando el árbol final.

Criterios de medida de impureza

Estas pruebas consisten en calcular la *optimalidad de cada atributo*, y se calcula a través de la *medida de ganancia de información* y *el índice de diversidad.* Estas medidas permiten tener una buena elección de la partición y al mismo tiempo aseguran no emplear un alto costo computacional. Estas medidas varían de acuerdo con cada algoritmo de árboles de decisión seleccionado.

El ***índice de diversidad,*** es la probabilidad de que el segundo atributo sea elegido dado que el primer atributo ya fue elegido y es distinto. Esto se conoce como la *probabilidad condicional*:

$$P\langle h \mid e \rangle = \frac{P(h \cap e)}{P(e)}; \text{donde } h, e \subseteq \Omega \text{ y } P(e) \geq 0$$

Así para un ejemplo binario, la posibilidad de elegir dos elementos iguales dentro de una población es $P(e) * P(e)$, y la posibilidad de elegir un elemento diferente en dos intentos es $1 - (P(e)^2 + P(e)^2)$

El valor máximo posible de *la diversidad* es de $(1/2)$, dado que hay solo dos clases, o $(1/n)$ donde n, es el número de clases y cada clase tiene el mismo número de miembros y por lo tanto la misma probabilidad de que sean elegidos.

Así, la fórmula de la diversidad para un índice binario es: $2p_1(1 - p_1)$.

Cuando no se le encuentra sentido a realizar una partición más al árbol, esto es, cuando la creación de un nodo no decrementa *el índice de la diversidad*, se deberá marcar como un nodo hoja, con lo que no se puede profundizar más.

¿Cuál debe ser la característica que primero particione el árbol?

Deberá ser aquella que marque *el índice de la diversidad* al 50%, con lo que se dividen dos grandes ramas del árbol. Para niveles inferiores, se deberán elegir los elementos que particionen al árbol de tal forma en el que una rama tenga la mayor probabilidad, es decir, que decremente la diversidad lo mayormente posible.

La ganancia de información, se utiliza para seleccionar los atributos de los nodos en cada prueba realizada y nos indica que el número de *bits* (partes) requeridos para describir una situación particular (resultado) depende del número de posibles salidas. Por ejemplo, si se tienen 8 clases igualmente probables, se obtiene el $\log_2(8) = 3$ bits, o si fueran 4 clases seria $\log_2(4) = 2$ bits. Para el caso de las 8 clases, si se logra particionar el conjunto en 4 clases, se dice que se tuvo una *ganancia de información* de 1 *bit* (parte).

Cada algoritmo tiene su propio criterio de partición, y su objetivo es conseguir nodos más puros, algunos criterios son: *el criterio del error esperado, el criterio Gini, los criterios Gain, Gain Ratio y la modificación del C4.5 y el DKM.*

Estos criterios de partición buscan la partición s con el menor $I(s)$, el cual se define de la siguiente forma:

$$I(s) = \sum_{J=1..n} p_j * f(p_{1j}, p_{2j}, ..., p_{cj})$$

Donde n es el número de nodos hijos de la partición (número de condiciones de la partición), pj es la probabilidad de "caer" en el nodo j, p_{1j} es la proporción d de elementos de la clase 1 en el nodo j, p_{2j} es la proporción de elementos de la clase 2 en el nodo j, y así para las c clases. Bajo esta fórmula general, cada *criterio de partición* implementa una función f distinta, como se muestra en la siguiente tabla:

Tabla 3.1 Criterio de partición

Criterio de partición o medidas de impureza	$f(p_{1j}, p_{2j},..., p_{cj})$
Error Esperado	$Min(p^1, p^2,..., p^c)$
GINI [26](CART)	$i(t) = \sum_{\substack{i,j=1 \\ i \neq j}}^{J} p(i\|t)p(j\|t) = 1 - \sum_{j=1}^{J} p(j\|t)^2$
Entropía [27](gain)	$i(t) = -\sum_{j=1}^{J} p(j\|t) \log p(j\|t)$ Se asume que 0 log 0=0
DKM	$2(\prod p_i)^{1/2}$

Las funciones $f(.)$ son funciones de impureza y, por tanto, la función $I(s)$ calcula la media ponderada (dependiendo de la cardinalidad de cada hijo) de la impureza de los hijos en una partición.

[26] El coeficiente de Gini es una medida de la desigualdad ideada por el estadístico italiano Corrado Gini. Normalmente se utiliza para medir la desigualdad en los ingresos, pero puede utilizarse para medir cualquier forma de distribución desigual. El índice de Gini es el coeficiente de Gini expresado en porcentaje, y es igual al coeficiente de Gini multiplicado por 100.

[27] El término Entropía significa evolución o transformación. En este caso el término se refiere al propuesto por Claude E. Shannon denominado Entropía en la Información, que expresa el grado de incertidumbre que existe sobre un conjunto de datos.

Ejemplo. Criterio de partición utilizando el índice de diversidad Gini

Se utilizan dos nodos muy diferentes, t_1 y t_2. A simple vista, parece que el nodo t_2 es más homogéneo o puro (presenta menos diversidad). Ver figura 3.12.

Figura 3.12 *Nodos para calcular* $i(t)$

De acuerdo a la expresión de la Tabla 3.1, el índice de diversidad Gini, queda de la siguiente manera:

$$i(t_1) = \sum_{\substack{i,j=1 \\ i \neq j}}^{3} p(i|t_1)\,p(j|t_1) = 1 - \sum_{j=1}^{3} p(j|t_1)^2$$

Ecuación 1 **Ecuación 2**

Según la ecuación 1:

$$i(t_1) = p(1|t_1)p(2|t_1) + p(1|t_1)p(3|t_1) + p(2|t_1)p(1|t_1) + p(2|t_1)p(3|t_1) + p(3|t_1)p(1|t_1) + p(3|t_1)p(2|t_1) =$$

$$= 2\,p(1|t_1)p(2|t1) + = 2\,p(2|t_1)p(3|t1) + 2\,p(3|t_1)p(1|t1) =$$

$$= 2\left(\frac{5}{40}\frac{10}{40}\right) + 2\left(\frac{10}{40}\frac{25}{40}\right) + 2\left(\frac{25}{40}\frac{5}{40}\right) = 0.5313$$

y según la ecuación 2: $i(t_1) = 1 - (p(1|t_1)^2 + p(2|t_1)^2 + p(3|t_1)^2) =$

$$= 1 - \left(\left(\frac{5}{40} \right)^2 + \left(\frac{10}{40} \right)^2 + \left(\frac{25}{40} \right)^2 \right) =$$

$$1 - \frac{25 + 100 + 625}{1600} = 1 - 0.4687 = 0.5313$$

La impureza del nodo t_2 será:

$$i(t_2) = 2p(1|t_2)p(2|t_2) + 2p(2|t_2)p(3|t_2) + 2p(3|t_2)p(1|t_2) =$$

$$= 2\left(\left(\frac{2}{40} \frac{3}{40} \right) + 2\left(\frac{3}{40} \frac{35}{40} \right) + 2\left(\frac{35}{40} \frac{2}{40} \right) \right) = 0.2263$$

lo que demuestra, numéricamente, que el nodo t_2 es menos impuro (más homogéneo) que el nodo t_1.

3.3.4 Regla de asignación de clases

¿Cómo asignar una etiqueta (clase) a un nodo terminal?

El objetivo es asignar una clase, j, a cada nodo terminal $t \in \tilde{T}$.

La clase asignada al nodo $t \in \tilde{T}$ se notará por $j(t)$. La forma más simple es la elección de la clase para la cual $p(j|t)$ es máxima *(la más representada en ese nodo)*.

$$j(t) = j \text{ si } p(j|t) = \max_{i=1,2,\ldots,J} \{p(i|t)\}$$

Si el máximo se alcanza para dos o más clases, se realiza un sorteo y se asigna arbitrariamente cualquiera de ellas.

3.3.5 La estrategia de poda

Los algoritmos de aprendizaje de *árboles de decisión,* obtienen modelos completos y consistentes con respecto a la evidencia. Estos modelos cubren todos los *ejemplos* de manera correcta. El *ajustarse demasiado* a la evidencia da como resultado que el modelo se comporte mal para nuevos *ejemplos*, ya que, en la mayoría de los casos, el modelo es solamente una aproximación del concepto objetivo del aprendizaje. Por tanto, intentar aproximar demasiado hace que el modelo sea demasiado específico, poco general y, por tanto, malo con datos no clasificados.

En segundo lugar, esto es especialmente patente cuando la evidencia puede contener ruido (errores en los atributos o incluso en las clases), ya que el modelo intentará ajustarse a los errores y esto perjudicará el comportamiento global del modelo aprendido, lo que se conoce como sobre-ajuste o sobre aprendizaje *(overfitting)*.

La manera más frecuente de limitar este problema es modificar los algoritmos de aprendizaje de tal manera que obtengan modelos más generales, lo que significa eliminar condiciones de las ramas del árbol o de algunas reglas. Este procedimiento se define como *poda.*

En la figura 3.13, se aprecia que los nodos que están por debajo del límite de *poda* se eliminan, ya que se consideran demasiado específicos.

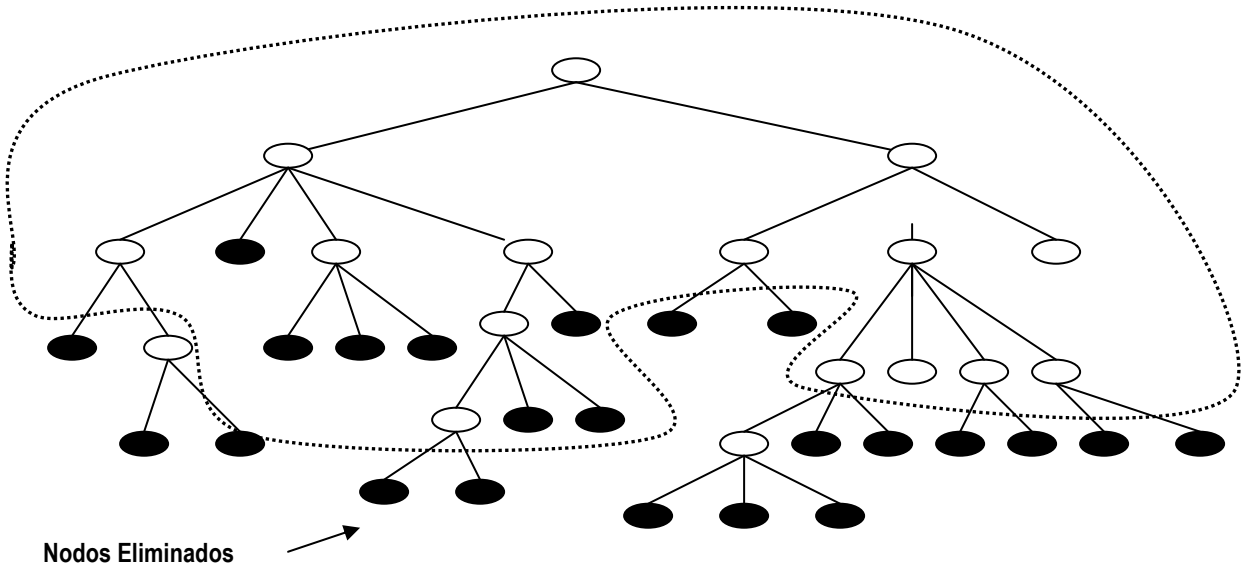


Figura 3.13 *Ejemplo del procedimiento de poda*

Definiendo el proceso de *poda* de una forma más precisa, se tiene que el nodo t, es el antecesor del nodo t', por lo tanto, en la figura 3.14 (a), $t_4, t_5, t_8, t_9, t_{10}$ y t_{11} son descendientes de t_2, no así t_6 y t_7. De forma similar, t_4, t_2 y t_1 son antecesores de t_9, pero t_3 no es antecesor de t_9.

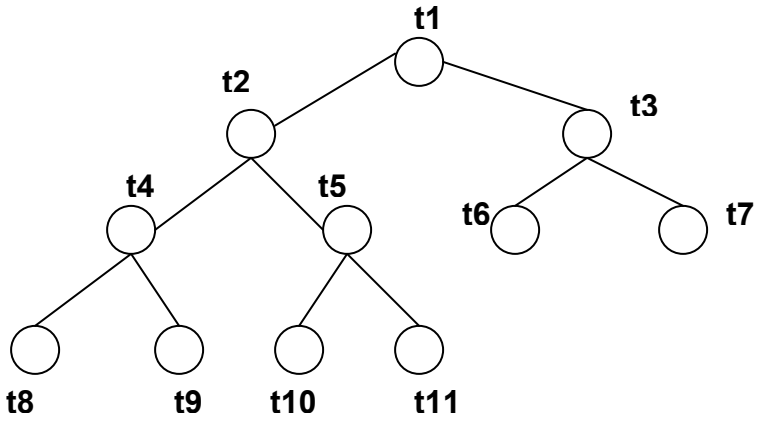


Figura 3.14 (a) *Árbol T.*

Definición: "Una rama T_t de T con nodo raíz $t \in T$ consiste del nodo t y todos los descendientes de t en T." (Breiman,1984). La rama T_{t2}, se ilustra en la figura 3.14 (b).

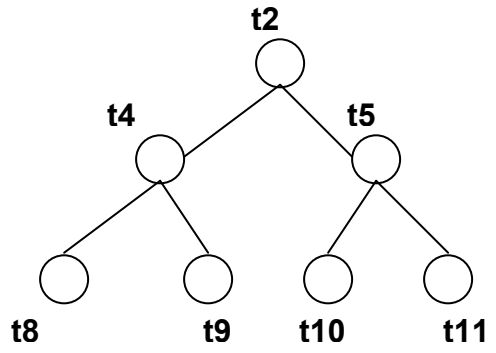


Figura 3.14 (b) *Rama* T_{t2}

Definición: "La poda de una rama T_t de un árbol T consiste en la eliminación de T de todos los descendientes de t, esto es, borrar todo T_t, exceptuando el nodo raíz. El árbol podado se denota por $T - T_t$." (Breiman,1984). El árbol podado $T - T_{t2}$, se ilustra en la figura 3.14 (c).

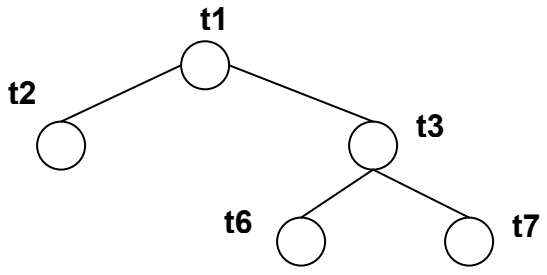


Figura 3.14 (c) *Árbol* $T - T_{t2}$

Definición: "Si T' se obtuvo de T por un procedimiento de poda consecutivo, entonces T' es llamado un subárbol podado de T y se denota por $T' < T$." (T' y T tienen el mismo nodo raíz)." (Breiman,1984).

Dentro del procedimiento de *poda*, se encuentran *la prepoda y la postpoda,* ambas técnicas pueden combinarse, una consecuencia de utilizar estos métodos, es que los nodos hoja ya no van a ser puros, es decir, que tengan ejemplos de varias clases.

- **Prepoda.-** Proceso que se realiza durante la construcción del árbol. Se determina el criterio de parada a la hora de seguir especializando una rama o una regla. En general, los criterios de *prepoda* pueden estar basados en el número de ejemplos por nodo, en el número de excepciones respecto a la clase mayoritaria *(error esperado)* o en técnicas más sofisticadas.

- **Postpoda.-** Proceso que se realiza después de la construcción del árbol. En los *árboles de decisión* se trata de eliminar nodos de abajo a arriba hasta un cierto límite.

Cada algoritmo tiene sus métodos para *podar árboles de decisión* y cada uno utiliza diferentes criterios.

Nivel óptimo de poda

La mayoría de métodos de *poda,* tienen parámetros para determinar al grado de *poda.* Aunque los algoritmos suelen tener un valor por defecto, es imposible encontrar un grado que funcione bien para todos los problemas. Existirán problemas con más ruido, con más *ejemplos*, o con características especiales que hagan necesario extremar o suavizar *la poda*.

Una manera de poder ajustar el nivel de *poda* a un determinado problema es observar el comportamiento con respecto a los datos de validación.

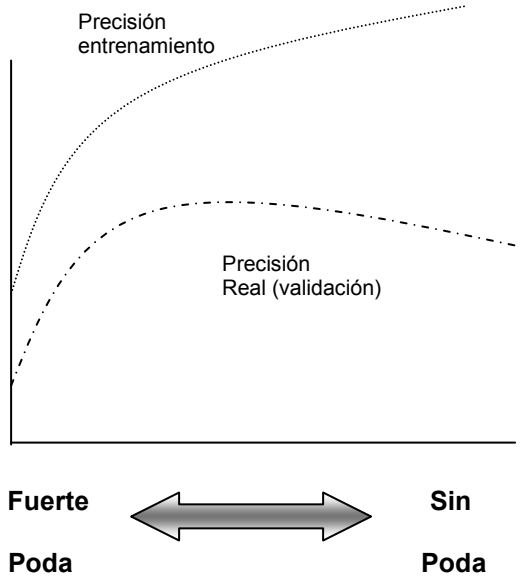


Figura 3.15 *El efecto beneficioso de la poda sólo es visible con datos con ruido y con los datos de validación*

La poda no es beneficiosa para los datos de entrenamiento, ya que los datos de entrenamiento no son precisamente una referencia objetiva. En cambio, el comportamiento respecto a los datos de validación es esclarecedor; para cada problema en concreto existe un grado de *poda* que es óptimo y que se puede estimar mediante el uso de estos datos de validación, si se dispone de ellos.

La *poda*, en conjunto con buenos *procedimientos de estimación (*método *cross-validation, incluido en CART v5.0)*, dan como resultado, *árboles de tamaño óptimo* y producen *estimaciones satisfactorias*.

Una de las características distintivas del algoritmo CART es su estrategia de poda, de hecho, Breiman (1984) incide en la importancia de la poda frente al de selección de particiones. Su argumento es que resulta más eficiente podar un árbol que detener su crecimiento: la poda permite que un subárbol de un nodo permanezca y el otro desaparezca, mientras que detener el crecimiento, poda *ambas* ramas simultáneamente.

En CART, el procedimiento general de poda, se resume como sigue:

1. Particionar nodos hasta que se cumpla alguna de estas condiciones:
   a) que sea totalmente puro, o
   b) $N(t) < N_{\min}$ (habitualmente $N_{\min} = 5$)

   Esto significa, que hasta que un nodo sea perfectamente homogéneo o hasta que tenga asociados a pocos *ejemplos.* Resulta evidente que el resultado debe ser un árbol muy grande, al que llamaremos $T_{\max}$.
2. Una vez obtenido $T_{\max}$ se trata de podar este árbol, obteniendo una secuencia *decreciente y anidada* de árboles.

   Si T'' se obtiene a partir de T por poda, T'' es un ***subárbol podado*** de T y se escribe $T'' \prec T.$ Así, la secuencia decreciente de árboles podados y anidados será la siguiente:

$$T_{\max} \succ T_1 \succ T_2 \succ \ldots \succ \{t_1\}$$

de manera que el árbol $\{t_1\}$ es un árbol que consta de un único nodo.

Uno de estos árboles será el que se seleccione finalmente:

Para realizar esta selección se asocia una medida de error a cada árbol de la secuencia y se escoge aquel que tenga asociado el menor error.

La idea básica es la de podar aquellos árboles que produzcan pequeños beneficios de bondad. Se espera que árboles podados (más simples) produzcan mejores resultados que los obtenidos con árboles más grandes (más complejos) al clasificar patrones independientes, esto es, los árboles podados tendrán más capacidad de generalización al no estar tan ajustados al conjunto de aprendizaje (el problema del sobre aprendizaje).

Poda por mínimo coste-complejidad

La medida de coste-complejidad involucra un proceso en el que un parámetro de penalización es continuamente incrementado.

La medida de coste-complejidad se define como sigue:

- Para cualquier subárbol $T \preceq T_{\max}$ se define su complejidad como el número de nodos terminales, $|\tilde{T}|$.
- El error de clasificación asociado al árbol T es $R(T)$.
- La medida de coste-complejidad asociada al árbol T, $R_\alpha(T) = R(T) + \alpha |\tilde{T}|$

donde α es un valor real $(\alpha \geq 0)$ (*parámetro de complejidad*) que se interpreta como el *coste de complejidad* por nodo terminal.

Así $R_\alpha(T)$ es una combinación lineal del coste del árbol y su complejidad, ponderada apropiadamente.

Un valor alto de α produce un incremento en el coste por nodos terminales, lo que se traduce en penalizar un subárbol con un alto número de nodos terminales.

Para cada α, se trata de encontrar el árbol $T(_{\alpha}), T(\alpha) \preceq T_{\max}$, que minimiza $R_{\alpha}(T)$,

$$R_{\alpha}(T(\alpha)) = \min_{T \preceq T_{\max}} \{R_{\alpha}(T)\}$$

Ejemplo. Funcionamiento del procedimiento de poda por coste-complejidad

Consideremos los siguientes árboles:

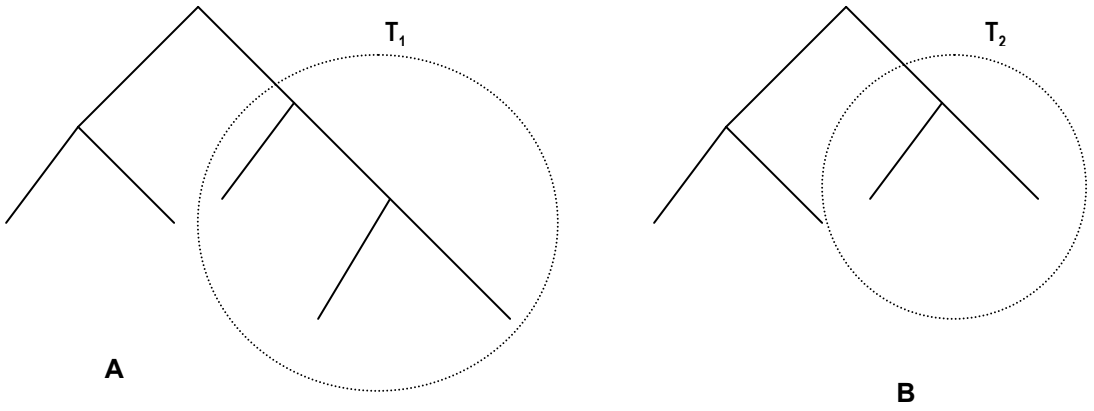


Figura 3.16 (a) *Subárbol* T_1. (b) *Subárbol* T_2

Se considera el caso de que el coste por nodo terminal sea 0.10, esto es, $\alpha = 0.10.$

1. Si nos centramos en el subárbol T_1 (figura 3.16 (a)) la complejidad de este árbol es $\left|\tilde{T}_1\right| = 3$.

Supongamos que $R(T_1) = 0.25.$ Entonces, la medida de coste-complejidad de T_1 es:

$$R_{\alpha}(T_1) = R(T_1) + \alpha\left|\tilde{T}_1\right| = 0.25 + (0.10 \times 3) = 0.55$$

2. Supongamos que se reemplaza el subárbol derecho de T_1 por una hoja, obteniendo T_2 (figura 3.16 (b)). En este caso, $|\tilde{T}_2|$=2. Este tipo de poda causa, generalmente, un incremento en el error, por lo que, por ejemplo, sea $R(T_2) = 0.38$.

$$R_\alpha(T_2) = R(T_2) + \alpha|\tilde{T}_2| = 0.38 + (0.10 \times 2) = 0.58$$

Para este valor de α, $R_\alpha(T_1) < R_\alpha(T_2)$, por lo que en este caso, el subárbol T_1 no se podaría.

Ahora consideremos el caso que $\alpha = 0.15$. Los valores de $R(T_1)$ y $R(T_2)$ se mantienen, por lo que las medias de coste-complejidad son:

$$R_\alpha(T_1) = R(T_1) + \alpha|\tilde{T}_1| = 0.25 + (0.15 \times 3) = 0.70$$

$$R_\alpha(T_2) = R(T_2) + \alpha|\tilde{T}_2| = 0.38 + (0.15 \times 2) = 0.68$$

En este caso, α, $R_\alpha(T_1) > R_\alpha(T_2)$, por lo que se prefiere T_2.

3.4 Árboles para regresión

La construcción de un *árbol de regresión* es muy similar a la tradicional para *clasificación,* pero con las siguientes diferencias:

- La función aprendida tiene dominio real (numérico) y no discreto (nominal).
- Los nodos hoja del árbol se etiquetan con valores reales, de tal manera que una cierta medida de calidad se maximice, por ejemplo la varianza de los *ejemplos* que caen en ese nodo respecto al valor asignado.

3.5 Árboles para agrupamiento o estimación de probabilidades

Además de su uso para *clasificación y regresión, los árboles de decisión* también han sido modificados para utilizarse en *agrupamiento.* La primera idea es modificar el criterio de partición y de evaluación para que considere particiones que separen entre zonas densas y poco densas.

Otra utilización de *los árboles de decisión* es *la estimación de probabilidades.* En este caso, la presentación del problema es similar a la de un problema de *clasificación;* los *ejemplos* tienen una etiqueta discreta, denominada clase. La diferencia es que se trata de determinar para cada nuevo *ejemplo* ¿cuál es la probabilidad de que pertenezca a cada una de las clases?

La modificación de un *árbol de decisión* para que sea *un estimador de probabilidades* es la siguiente:

Supongamos que se tienen tres clases: $a,\ b\ y\ c$. Para cada nodo hoja con una cardinalidad n, tendremos un determinado número de *ejemplos* de cada clase: $n_a, n_b\ y\ n_c$. Si dividimos cada uno de estos valores por la cardinalidad total, tendremos una estimación de las probabilidades de las clases en es nodo, es decir: $p_a = n_a/n,\ p_b = n_b/n\ y\ p_c = n_c/n$. Este tipo de árboles de decisión modificados de esta manera se conocen como PETs *(Probability Estimation Trees)*.

3.6 Árboles para grandes volúmenes de datos

Los algoritmos de aprendizaje de decisión, debido a su carácter voraz y a su estructura de “divide y vencerás”, se comportan especialmente bien con grandes volúmenes de datos, ya sean de *gran dimensionalidad* (muchos atributos) o de *gran cardinalidad* (muchos ejemplos). Este buen comportamiento (demostrado teórica y experimentalmente) muchas veces se basa en asumir que los datos caben en memoria, si los datos no caben en memoria, el rendimiento de los algoritmos se degradará por la constante consulta de información de disco a memoria.

En la última década se han modificado los algoritmos para que sean más eficientes en el contexto de grandes tablas, intentando minimizar los barridos sobre tablas y el uso de accesos a disco. En general, los algoritmos llamados “algoritmos de aprendizaje de árboles de decisión escalables”, se vuelven a diseñar teniendo en cuenta varios requisitos.

En primer lugar, no requieren que los datos estén en memoria y en segundo lugar, las comprobaciones de consistencia de las condiciones se deben hacer eficientemente, utilizando índices, con el objetivo de agilizar los barridos de los datos.

3.7 Algoritmos

Los algoritmos de *árboles de decisión* se clasifican en dos grandes grupos: *los árboles clasificadores para el aprendizaje (Machine learning) y reconocimiento de patrones (Pattern recognition) con volúmenes de datos estándar,* tales como ID3, C4.5, CART, CHAID y *los clasificadores enfocados a grandes volúmenes de datos y gestión de memoria secundaria*, como son: SONAR, RAINFOREST, SLIQ y SPRINT.

Cada uno de los anteriores utilizan diferentes criterios de partición y llegan a obtener resultados ligeramente diferentes.

- **CHAID (*CHI-squared Automatic Interaction Detector*)**, es descendiente del *Sistema de Detección de Interacción Automática (AID).* Realiza particiones no binarias y usa *test-f* para determinar la partición óptima en el caso de *regresión* y *test ji-cuadrado* para *clasificación* para crear múltiples divisiones. Es el más ampliamente utilizado en paquetes estadísticos como *(SPSS y SAS)* y antecede, y requiere más preparación de datos que CART.

- **ID3, C4.5 y derivados (Assistant**): son métodos “divide y vencerás” y están basados en criterios de partición derivados de la ganacia *GainRatio*. Tiene estrategia de poda basada en reglas u otros mecanismos más sofisticados. Contiene métodos de colapsado de ramas y muchas otras mejoras. Una versión más avanzada es la J4.8, y existe también la versión C5.

- **IND, LMDT** y otros sistemas híbridos, incorporan características de varios sistemas o añaden técnicas de aprendizaje en la construcción de *árboles de decisión: regresión lineal, perceptrones, etc.*

- **SLIQ, SPRINT, SONAR y RAINFOREST,** son modificaciones de los árboles de decisión clásicos para conseguir escalabilidad para grandes volúmenes de datos, paralelización, etc.

3.8 Ventajas y desventajas de los árboles de decisión

Las principales ventajas que representan *los árboles de decisión* son:

1. Generan modelos comprensibles e inteligibles y proposicionales para la toma de decisiones del negocio, ya que expresa de forma simbólica un conjunto de reglas, fáciles de "traducir" al idioma humano.

2. Son aplicables a varias tareas de minería de datos*: clasificación, agrupamiento y estimación de probabilidades.*

3. Tratan con atributos numéricos (continuos) y nominales (discretos).

4. Muchos de ellos son eficientes y existen variantes escalables a grandes volúmenes de datos (tanto para muchos atributos como muchos *ejemplos*).

5. Son tolerantes al ruido, a atributos no significativos y a valores faltantes.

6. Mejoran el rendimiento de *la clasificación* sin requerir mucho poder de cómputo.

7. Indican claramente ¿cuál de los campos es más importante para una *clasificación o predicción*?, esto es porque los campos patrocinadores más cercanos al nodo raíz, son los campos más característicos para el estudio realizado, es decir, son los más determinantes.

8. En general, los pasos de clasificación y de aprendizaje en *los árboles de decisión* son generalmente rápidos.

9.Existe *software* de distribución gratuita.

Algunas desventajas son:

1. No son tan precisos como otros métodos, como pueden ser *las redes neuronales o las máquinas de vectores de soporte.*

2. Son aprendices débiles *(weak learners)*, debido a su carácter voraz, son bastante dependientes de la muestra de *ejemplos*; dos muestras distintas sobre la misma distribución pueden dar dos árboles bastante diferentes.

3. Son problemáticos para datos de *series de tiempo,* a menos que un gran esfuerzo sea puesto en la preparación de datos, de tal forma que las tendencias y patrones sean visibles.

4. *Los árboles de decisión* son menos apropiados para tareas de *estimación* donde la meta es predecir el valor de variables continuas.

5. Existen restricciones en algunos algoritmos, ya que sólo pueden manejar valores binarios (sí/no, aceptar/rechazar), pero para los casos en los que pueden obtenerse muchos nodos hijos, existe una tendencia a obtener una tasa de error alta por el crecimiento desmesurado del árbol, costando muchos recursos de cómputo.

Capítulo 4

ANÁLISIS ESTADÍSTICO DESCRIPTIVO

Palabras Clave: Población. Muestra. Variables Categóricas. Variables Cuantitativas. Escala.

4.1 Antecedentes

En este capítulo se desarrollan las dos primeras fases del proceso KDD (integración y recopilación de información y la de selección, limpieza y transformación de los datos). Es de importancia señalar que para la investigación no se contó con gran volumen de información histórica que pudiera ser utilizada para conocer el comportamiento de la población, por lo que fue necesario recavar la información para formar el almacén. También es importante señalar, que si bien un almacén de datos es muy aconsejable para la minería de datos, no es imprescindible y en algunos casos, cuando el volumen de información no es muy grande, se puede trabajar con los datos originales o en formatos heterogéneos (archivos de texto, hojas de cálculo).

Primeramente se definió la población bajo estudio, posteriormente se obtuvo una muestra de la misma y con ello se realizó el análisis estadístico descriptivo para conocer sus características.

4.2 Definición de la población

Recordemos que la fuerza de ventas de la compañía K se forma por las personas que están en los diferentes niveles de distribuidor y de líderes de venta.

El objetivo de definir a la población bajo estudio es el "identificar a las personas de la compañía K, que formarán parte del grupo de estudio".

Para determinar los elementos que se incluirán se utilizaron los siguientes criterios de exclusión:

- Las personas deben pertenecer a una la línea de auspicio mexicana.
- Se tomaron a todas las personas que actualmente son líderes de venta.
- Las actuales líderes son de nacionalidad mexicana y su lugar de residencia es el Distrito Federal.
- Las líderes no deben tener vínculos familiares con otras líderes de venta o con personal de la compañía.
- Las líderes deben estar activas.[28]

[28] *Activa,* significa que la Líder realiza con regularidad órdenes de compra a la compañía, al igual que su grupo de compra.

De las 30 líneas de auspicio muy exitosas que tiene la compañía K en México, la que otorgó las facilidades para este estudio, cuenta hoy en día con una población de 383 líderes de venta.

Determinación de la muestra

Aplicando los criterios anteriores, resulta que de la población total de 383 líderes, 120 líderes no viven en el D.F. y las restantes 63 no están activas, por lo que se trabajó con una muestra de 200 elementos. Esta muestra se eligió a juicio[29], es decir, en conjunto con la líder de la línea de auspicio que facilitó este trabajo.

En la muestra se identificaron a las líderes que obtuvieron el liderazgo en un año denominándolas el grupo exitoso con 80 elementos y aquellas que lo obtuvieron en un tiempo mayor conocidas como el grupo no exitoso, conformado por 120 personas.

4.3 Integración y recopilación de la información

Para recopilar la información se elaboró y aplicó un cuestionario de reconocimiento[30] con la finalidad de "detectar las características de comportamiento que influyen en las mujeres, para obtener el nivel de líder de ventas dentro de la compañía K". El cuestionario se aplicó a los dos grupos (exitoso y no exitoso) y se solicitó a las actuales líderes enfocarse al "tiempo pasado", esto es, cuando tenían el nivel de

[29] Cuando se obtiene una muestra de juicio, la persona que elabora la muestra elige unidades que considera representativas de la población. La validez de una muestra de juicio refleja la solidez del juicio del recolector de datos.

[30] El cuestionario se llama de "reconocimiento" por que está orientado a identificar características personales de las líderes. Cabe aclarar que para la elaboración del cuestionario no se consultó a ningún profesional, sino que tanto su elaboración y aplicación fue supervisada y asesorada por líderes de venta del grupo exitoso.
El cuestionario se encuentra en el Apéndice A.

distribuidoras, para conocer su situación y evaluar las causas que las llevaron a este nivel.

Antes de presentar las variables, su medición y su escala, se definirán estos conceptos.

Variables categóricas.- Son aquellas en las que el estado de su respuesta está dado por dos o más eventos. **Ejemplo:** la probabilidad de compra de un producto o el número de ofertas de vacaciones aceptadas o el tiempo que lleva un cliente sin comprar por Internet.

Medición.- Asignación de números u otros signos a las características de los objetos de acuerdo con ciertas reglas especificadas previamente.

Escala.- Es una extensión de la medición. Las escalas comprenden la creación de un continuo sobre el que se localizan los objetos medidos. Las escalas de medición que se utilizan en este trabajo son: nominal, ordinal, de intervalos, de relación y de Likert. Ver Tabla 4.1.

Para la aplicación del cuestionario de reconocimiento, se acudió personalmente con cada líder y se le dio el tiempo suficiente para contestar, informándole el grado de confidencialidad del mismo, y para evitar los datos faltantes o erróneos, la revisión de cada cuestionario se llevó a cabo en el mismo momento.

En la elaboración del cuestionario, se definieron dieciocho variables iniciales, así como su tipo, su escala de medición, su categoría y su codificación. (Ver Tabla 4.2).

Tabla 4.1.Escalas de medición

Escala	Características	Ejemplos	Ejemplos mercadológicos	Estadísticas permitidas	
				Descriptivas	**Deductivas**
Nominal	Los números identifican y clasifican los objetos.	Números de afiliación al seguro social, números de jugadores de fútbol.	Números de marcas, tipos de tiendas, clasificación por sexo.	Porcentajes, modo.	Chi-cuadrada, prueba nominal.
Ordinal	Los números indican las posiciones relativas de los objetos, pero no la magnitud de las diferencias entre éstos.	Clasificaciones de calidad, clasificaciones de los equipos en un torneo.	Clasificaciones por preferencias, posición en el mercado, clase social.	Percentilio, mediana.	Correlación, ANOVA.
De intervalos	Las diferencias entre los objetos pueden compararse; el punto cero es arbitrario.	Temperatura	Actitudes, opiniones, números en un índice.	Rango, media, desviación estándar	Correlaciones, producto-momento, pruebas *t,* ANOVA, regresión, análisis de factores.
De relación	El punto cero es fijo, las relaciones de los valores de la escala pueden calcularse.	Longitud, peso	Edad, ingreso, costos, ventas, participación en el mercado.	Media geométrica, media armónica.	Coeficiente de variación.
Likert	Escala de clasificación, en que a los entrevistados se les pide indiquen un grado de acuerdo o desacuerdo con cada una de la serie de afirmaciones respecto a los objetos de estímulo.	Identificar en el rango de: -Por completo en desacuerdo -En desacuerdo -Neutral -De acuerdo -Por completo de acuerdo.			

4.3.1 Definición y descripción de variables iniciales

Tabla 4.2 Definición y descripción de las variables iniciales

Variable	Tipo de Variable	Escala	Código / Categoría[31]
T_CIA: Tiempo que lleva la líder de ventas en la compañía K	Cuantitativa	Continua	1,2,3,... años
T_LIDER: Tiempo de ser líder de ventas	Cuantitativa	Continua	1,2,3,... años
HIJOS: Evaluación de los hijos que son dependientes económicos de la líder	Categórica dicotómica	Nominal	0 = No, tiene hijos dependientes 1 = Sí, tiene hijos dependientes
FACT_INF: Factores positivos o negativos que influyeron en la toma de decisión	Categórica politómica	Nominal	1 = Necesidad económica 2 = Independencia laboral 3 = Confianza en la compañía 4 = Los hijos 5 = Autorrealización 6 = Un cambio de vida 7 = Reconocimiento profesional 8 = Ser una ejecutiva independiente 9 = Timidez 10 = Pareja en desacuerdo 11 = Interés único por las ventas 12 = No había quien cuidara a hijos 13 = Trabajo adicional absorbente 14 = Desconocimiento de beneficios 15 = Sin experiencia 16 = Falta de decisión 17 = Falta de disciplina 18 = Falta de capacitación
TOMAR_DE: Evaluación de la toma de decisión para ser líder en un año	Categórica dicotómica	Nominal	1 = Sí, le fue difícil tomar la decisión. 2 = No, le fue difícil tomar la decisión.

[31] La designación de las categorías fue realizada con base a la experiencia de las líderes de venta.

Variable	Tipo de Variable	Escala	Código / Categoría
PERS_INF: Persona que influyó determinantemente en la toma de decisión	Categórica politómica	Nominal	1 = Líder del grupo 2 = Hijos 3 = Familia 4 = Ella misma 5 = Otra
EXP_VENT: Evaluación de las experiencia en ventas	Categórica dicotómica	Nominal	1 = Sí, tiene experiencia 2 = No, tiene experiencia
EMPLEO_2: Evaluación del empleo adicional de la líder	Categórica politómica	Nominal	1 = Profesional o Técnica 2 = Gerentes o Administradoras 3 = Ama de Casa 4 = Vendedoras 5 = Trabajadoras de oficina 6 = Artesanas / Obreras / Limpieza 7 = Sector educativo 8 = Otra
MOTDEMP: Motivo principal para dejar el empleo adicional	Categórica politómica	Nominal	1= Mayor remuneración económica. 2 = Mayor reconocimiento 3 = Mayor capacitación 4 = Horario flexible 5 = Independencia laboral
PRIORNEG: Prioridad del negocio para la líder	Categórica politómica	Ordinal	1 = Alta 2 = Media 3 = Baja
EDAD: Edad de la líder cuando alcanzó este nivel	Cuantitativa	De Relación	1,2,3,.....años
ESCOLAR: Grado máximo de escolaridad de la líder cuando alcanzó este nivel	Categórica politómica	Ordinal	1 = Ninguna 2 = Primaria 3 = Secundaria 4 = Técnico/Comercio 5 = Bachillerato 6 = Licenciatura 7 = Postgrado
EDO_CIV: Estado civil de la líder cuando alcanzó este nivel	Categórica politómica	Nominal	1 = Soltera 2 = Casada 3 = Viuda 4 = Divorciada / Separada 5 = Unión Libre

Variable	Tipo de Variable	Escala	Código / Categoría
GUSTCIA: Factor de la compañía K de mayor agrado para la líder	Categórica politómica	Nominal	1 = La filosofía y misión de la empresa 2 = Ganancia económicas 3 = Gente 4 = Productos 5 = Reconocimientos 6 = Negocio Independiente
EXPCIA:. Evaluación de las expectativas de la líder sobre la empresa K	Categórica dicotómica	Nominal	1 = Siempre 2 = Casi siempre
NOLGPO: Número actual de líderes desarrolladas en el grupo de compra de la entrevistada.	Cuantitativa	Continua	0,1,2,...personas
FACT_IMP: Factor principal que influye para no obtener el nivel	Categórica politómica	Nominal	1 = Miedo 2 = Hijos 3 = Pareja 4 = Otro empleo 5 = Desconocimiento o Ignorancia 6 = Falta de Capacitación 7 = Impedimento físico
ESTLIDER: Número estimado de líderes, a desarrollar en el grupo de compra por año	Cuantitativa	Continua	1,2,3,...personas

Una vez aplicado el cuestionario, se vació la información codificada a los archivos correspondientes para analizar el comportamiento de la muestra.

4.4 Análisis descriptivo de la muestra

4.4.1 Análisis de las variables categóricas

Al realizar el análisis de las trece variables cualitativas, involucradas en la obtención del liderazgo en un año, se obtuvo:

1) Hijos dependientes económicos

La Tabla 4.3 muestra para el grupo exitoso que el 82% de las mujeres exitosas tienen hijos que dependen económicamente de ellas, esto puede deberse a que están en edad escolar. El comportamiento en ambos grupos es muy similar.

Tabla 4.3 Número de casos para la variable HIJOS

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=20***	
Hijos dependientes económicos	**Casos**	**% casos**	**Casos**	**% casos**
Sí	66	82	106	88
No	14	18	14	12

2) Factores positivos o negativos que influyeron en la toma de decisión

En la tabla 4.4 se observa que en el grupo exitoso los factores que más influyeron en las mujeres para llegar al liderazgo son:

- **los hijos con 23%**
- la necesidad económica y la independencia laboral con 19%
- el ser una ejecutiva independiente con 12% y
- la autorrealización con el 10%.

Mientras que en el grupo no exitoso, los factores de mayor incidencia son:

- **la falta de decisión con un 15%**
- el trabajo adicional absorbente con un 13%
- la necesidad económica con un 13%
- el interés único por las ventas con un 10%

Tabla 4.4. Número de casos para la variable FACT_INF

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Factores	**Casos**	**% casos**	**Casos**	**% casos**
Necesidad económica	15	19	15	13
Independencia laboral	15	19	8	7
Confianza en la compañía	4	5	1	1
Los hijos	19	23	11	9
Autorrealización	8	10	5	4
Cambio de vida	3	3	6	5
Reconocimiento profesional	6	8	4	3
Ejecutiva independiente	9	12	6	5
Timidez	0	0	6	5
Pareja en desacuerdo	0	0	4	3
Interés único por ventas	0	0	12	10
No tiene quien cuide a los hijos	0	0	6	5
Trabajo adicional absorbente	0	0	16	13
Desconocimiento de beneficios	1	1	1	1
Sin experiencia	0	0	0	0
Falta de decisión	0	0	18	15
Falta de disciplina	0	0	1	1
Falta de capacitación	0	0	0	0

3) Evaluación de la toma de decisión para ser líder en un año

La tabla 4.5 presenta el número de casos y los porcentajes para la evaluación ante la toma de decisión de ser líder de ventas.

De forma notoria se observa que en el grupo exitoso al 72% no le costó trabajo decidirse y en el otro grupo al mismo porcentaje de la muestra sí le costó trabajo, este resultado se confirma con lo obtenido en el punto anterior, donde el factor de falta de decisión ocupa el mayor porcentaje.

Tabla 4.5 Número de casos para la variable TOMAR_DE

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Evaluación de la toma de decisión	**Casos**	**% casos**	**Casos**	**% casos**
Sí	22	28	86	72
No	58	72	34	28

4) Persona que influyó determinantemente en la toma de decisión

En la tabla 4.6 se observa que las personas que mayor influencia han tenido para las mujeres que lograron el liderazgo en un año son:

- **27% su familia**
- 25% sus hijos
- 23 % ella misma

Mientras que las mujeres que obtuvieron el liderazgo en un tiempo mayor, su principal influencia fue:

- **37% su líder de grupo**
- 23% sus hijos y familia
- 11% ella misma

Tabla 4.6 Número de casos para la variable PERS_INF

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Personas que influyeron	**Casos**	**% casos**	**Casos**	**% casos**
Líder del grupo	15	19	44	37
Hijos	20	25	28	23
Familia	22	27	27	23
Ella misma	18	23	14	11
Otra	5	6	7	6

5) Evaluación de la experiencia en ventas

En cuanto a la experiencia en el área de ventas, en el grupo exitoso el 57% de los casos contaba con experiencia en ventas, mientras que el otro grupo el 18% no la tenía. Ver Tabla 4.7.

Tabla 4.7 Número de casos para la variable EXP_VENT

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Experiencia en ventas	**Casos**	**% casos**	**Casos**	**% casos**
Sí	46	57	22	18
No	34	43	98	82

6) Empleo anterior de la líder

Se observa en la Tabla 4.8, que la ocupación de las mujeres exitosas era:

- **29% en el sector educativo**
- 20% profesionistas o áreas técnicas
- 20% trabajaban como gerentes o administradoras

En el segundo grupo su principal ocupación era:

- **27% ama de casa**
- 23% trabajaban en oficina
- 15% profesionistas o áreas técnicas

Tabla 4.8 Número de casos para la variable EMPLEO_2

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Ocupación anterior	**Casos**	**% casos**	**Casos**	**% casos**
Profesional / Técnica	16	20	18	15
Gerentes o Administradoras	16	20	5	4
Ama de casa	7	9	32	27
Vendedoras	6	7	14	11
Trabajadora de oficina	11	14	27	23
Artesana / Obrera / Limpieza	1	1	12	10
Sector educativo	23	29	12	10
Otro	0	0	0	0

7) Motivo principal para dejar su empleo anterior

Se observa en la Tabla 4.9 que en ambos grupos, las actuales líderes dejaron su ocupación anterior para recibir mayor remuneración económica en su negocio

independiente de venta directa con el 46%, seguido por el reconocimiento a su trabajo y la independencia laboral.

Tabla 4.9 Número de casos para la variable MOTDEMP

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Motivo para dejar empleo	**Casos**	**% casos**	**Casos**	**% casos**
Mayor remuneración económica	36	46	54	46
Mayor reconocimiento	13	16	35	29
Mayor capacitación	1	1	0	0
Horario flexible	9	11	15	12
Independencia laboral	21	26	16	13

8) Prioridad del negocio para la líder

La Tabla 4.10 muestra que para el grupo exitoso la prioridad de negocio es alta con el 78% comparado con el 36% del grupo no exitoso.

Tabla 4.10 Número de casos para la variable PRIORNEG

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Prioridad	**Casos**	**% casos**	**Casos**	**% casos**
Alta	62	78	43	36
Media	18	22	60	50
Baja	0	0	17	14

9) Grado máximo de escolaridad

La Tabla 4.11 muestra que en el grupo exitoso su nivel de escolaridad es más elevado, ya que el 44% estudió una licenciatura, seguido por el bachillerato con el 23%.

En el segundo grupo, la mayoría de las mujeres estudiaron el bachillerato o una carrera técnica, siendo el 34%.

Tabla 4.11 Número de casos para la variable ESCOLAR

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Escolaridad	**Casos**	**% casos**	**Casos**	**% casos**
Ninguna	1	1	0	0
Primaria	1	1	5	4
Secundaria	3	4	16	13
Técnico / Comercio	16	20	41	34
Bachillerato	18	23	41	34
Licenciatura	35	44	15	13
Postgrado	6	7	2	2

10) Estado civil

La Tabla 4.12, se observa que en el grupo exitoso:

- **54% son casadas**
- 43% están solteras

En el grupo no exitoso se tiene que:

- **50% son casadas**
- 19% están divorciadas o separada

Tabla 4.12 Número de casos para la variable EDO_CIV

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Estado civil	**Casos**	**% casos**	**Casos**	**% casos**
Soltera	34	43	18	15
Casada	43	54	60	50
Viuda	2	2	7	6
Divorciada / Separada	1	1	23	19
Unión libre	0	0	12	10

11) Factor de la compañía K de mayor agrado para la líder

Se observa en la Tabla 4.13 que para el grupo exitoso, los factores que más le agradan de la compañía K son:

- **Al 36% les agrada la filosofía o misión de la compañía**
- Al 24% les gusta el reconocimiento por su trabajo
- Al 14% le gustan las ganancias económicas

El grupo no exitoso opinó lo siguiente:

- **El 28% opinó que le agradan las ganancias económicas**
- El 24% le gusta el reconocimiento por su trabajo
- El 23% les agrada la filosofía o misión de la compañía

Tabla 4.13 Número de casos para la variable GUSTCIA

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Factor de mayor agrado	**Casos**	**% casos**	**Casos**	**% casos**
Filosofía o misión de la compañía	29	36	27	23
Ganancias económicas	11	14	34	28
Gente	5	6	7	6
Productos	1	1	8	7
Reconocimientos	19	24	29	24
Negocio Independiente	15	19	15	12

12) Evaluación de las expectativas de la líder sobre la empresa K

En la Tabla 4.14 se observa que al 64% de los casos del grupo exitoso sus expectativas sobre la compañía siempre han sido satisfechas.

En el segundo grupo el 46% de los casos manifestaron que sus expectativas casi siempre han sido cubiertas.

Tabla 4.14 Número de casos para la variable EXPCIA

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Expectativas	**Casos**	**% casos**	**Casos**	**% casos**
Siempre	51	64	55	46
Casi siempre	29	36	65	54

13) Factor principal que influye para que no se llegue al liderazgo

De acuerdo con las entrevistadas del grupo exitoso el 61% opinó que las excusas principales de las nuevas distribuidoras para no alcanzar este nivel son el miedo de afrontar nuevas responsabilidades y los hijos en edad escolar, de la misma manera el otro grupo expresó que el miedo es un factor que viene ligado al desconocimiento o ignorancia de los beneficios que obtendrían si alcanzan el nivel. Ver Tabla 4.15.

Tabla 4.15 Número de casos para la variable FACT_IMP

Categoría	**Exitoso** ***n=80***		**No exitoso** ***n=120***	
Factor de mayor agrado	**Casos**	**% casos**	**Casos**	**% casos**
Miedo	49	61	61	51
Hijos	1	1	11	9
Pareja	4	5	12	9
Otro empleo o negocio	4	5	8	7
Desconocimiento	22	28	26	22
Falta de capacitación	0	0	2	2
Impedimento físico	0	0	0	0

En resumen, las mujeres que lograron el liderazgo en un año, tienen las siguientes características, (ver Tabla 4.16):

Mujeres casadas, con hijos en edad escolar, los cuales fueron el motivo para incrementar su nivel en la compañía y obtener mayor remuneración económica. No les fue difícil tomar la decisión de llegar a este nivel, ya que su familia es la principal motivación e influencia para sus proyectos. Estas mujeres ya tenían experiencia en ventas con negocios de este tipo ya sea por su cuenta o en otras compañías, además de ser profesionistas a nivel licenciatura. Su ocupación anterior de la mayoría de las mujeres era como profesoras de grupo. Para ellas su negocio independiente en la compañía K siempre ha tenido alta prioridad, así mismo la filosofía y misión de la compañía las impulsa a seguir dentro de ella y sus expectativas siempre han sido cumplidas. Ellas opinan que las excusas de la mayoría de las distribuidoras para no obtener este nivel son el miedo y los hijos en edad escolar.

Las mujeres que por el contrario lograron el liderazgo, pero tardaron más tiempo en obtenerlo tienen las siguientes características:

Mujeres casadas con hijos en edad escolar. Son mujeres que para tomar la decisión de ser líderes les costo trabajo, pero fueron impulsadas y motivadas por su líder de grupo.

La mayoría eran amas de casa con estudios de bachillerato, que deseaban obtener mayor remuneración económica por su trabajo, pero sin descuidar su familia, por ello su negocio tenía una prioridad media. Lo que más le atraía de la compañía eran las ganancias económicas y los premios, sus expectativas de la compañía casi siempre se han cumplido. Ellas opinan que el miedo a afrontar nuevas responsabilidades frena a las nuevas distribuidoras para llegar a este nivel.

Tabla 4.16 Resumen de características de ambos grupos

Variables Categóricas	**Grupo exitoso**	**Grupo No exitoso**
Hijos dependientes económicos	82%	88%
Factores positivos o negativos que influyeron en la toma de decisión	23% fueron motivadas por sus hijos	15% la falta de decisión, hizo que obtuviera el liderazgo en mayor tiempo
Evaluación de la toma de decisión para ser líder en un año	72% no le costó trabajo	72% le costó trabajo
Persona que influyó determinantemente en la toma de decisión	27% fue motivada por su familia	37% su líder de grupo la impulsó a lograr este nivel
Evaluación de la experiencia en ventas	57% tenía experiencia	82% no la tenía
Empleo anterior de la líder	29% se desarrollaban como profesoras	27% eran amas de casa
Motivo principal para dejar el empleo adicional	46% mayor remuneración económica	46% mayor remuneración económica
Prioridad del negocio para la líder	78% prioridad alta	50% prioridad media
Grado máximo de escolaridad	44% licenciatura	34% bachillerato
Estado civil	54% casadas	50% casadas
Factor de la compañía K de mayor agrado para la líder	36% filosofía y misión	28% ganancias económicas
Evaluación de las expectativas de la líder sobre la empresa K	64% siempre cubiertas	54% casi siempre cubiertas
Factor principal que influye para que no se llegue al liderazgo	61% miedo a nuevas responsabilidades	51% miedo a nuevas responsabilidades

4.4.2 Análisis de las variables cuantitativas

Al realizar el análisis de las cinco variables cuantitativas, se obtuvieron los estadísticos básicos como la media, mediana, moda y desviación estándar:

1) Tiempo que lleva la líder en la compañía K

Tabla 4.17 Estadísticos la variable T_CIA

Estadísticos	**Grupo exitoso N=80**	**Grupo No exitoso N=120**
Media	4.00	5.37
Mediana	4.00	5.00
Moda	2	5
Desviación estándar	1.86	1.81

En la Tabla 4.17 se observa que, la media de tiempo que tienen las mujeres de permanencia en la compañía en el grupo exitoso es de 4 años, teniendo 36 casos por debajo de la media y 30 casos por arriba de ésta. No existe gran variabilidad o dispersión en los datos, ya que la desviación es de 1.86. Sin embargo, el número de años que más frecuencia tiene en el grupo es de 2 años en la compañía. La mediana es de 4 años.

En el grupo no exitoso, la media de tiempo es mayor que en el anterior por 1.37 años, teniendo 40 casos por abajo y 55 por arriba de ésta. Al igual que el anterior no existe gran variabilidad en los datos y el tiempo de mayor frecuencia en la compañía es de 5 años.

2) Tiempo de ser líder de ventas en la compañía K

Tabla 4.18 Número de casos para la variable T_LIDER

Estadísticos	**Grupo exitoso N=80**	**Grupo No exitoso N=120**
Media	3	2.35
Mediana	3.00	2.00
Moda	1	2
Desviación estándar	1.88	1.46

La Tabla 4.18 muestra que, la media de tiempo que tienen las mujeres de ser líderes en el grupo exitoso es de 3 años, teniendo 39 casos por debajo de la media y 30 casos por arriba de ésta. No existe gran variabilidad o dispersión en los datos, ya que la desviación es de 1.88. En el otro grupo la media de tiempo en ser líderes es de 2.35 años, 55 casos por abajo de ésta y 20 casos por arriba. No existe gran variabilidad o dispersión en los datos, ya que la desviación es de 1.46.

3) Edad

Tabla 4.19 Número de casos para la variable EDAD

Estadísticos	**Grupo exitoso N=80**	**Grupo No exitoso N=120**
Media	35.48	41.70
Mediana	34.50	44.00
Moda	45	52
Desviación estándar	6.82	12.03

La media de edad en el grupo exitoso es de 35.48 años, con 40 casos por abajo y 30 casos por arriba. En el otro grupo la edad promedio es de 41.70 años, con 53 casos por abajo y 63 casos por arriba. La edad con más frecuencia es de 45 años en el primer grupo y en el segundo es de 44 años. En ambos grupos se encuentran rangos de edad

muy diversos, acentuándose mayormente en el grupo no exitoso con una desviación estándar de 12.03

4) Número actual de líderes desarrolladas en el grupo de compra de la entrevistada

Tabla 4.20 Estadísticos para la variable NOLGPO

Estadísticos	**Grupo exitoso N=80**	**Grupo No exitoso N=120**
Media	1.14	0.70
Mediana	1.00	0.00
Moda	1	0
Desviación estándar	1.28	0.98

En la Tabla 4.20 se muestra de manera notoria que el grupo exitoso ha desarrollado en su grupo de compra, una líder de ventas en promedio, algunas mujeres de este grupo tienen hasta 2 líderes en su grupo, ya que la desviación estándar es de 1.28. El otro grupo todavía no ha logrado desarrollar a nuevas líderes en sus grupos, y en el mejor de los casos, hay mujeres que están casi por obtenerlo.

5) Número de líderes de venta que le gustaría desarrollar en su grupo de compra por año

Tabla 4.21 Estadísticos para la variable ESTLIDER

Estadísticos	**Grupo exitoso N=80**	**Grupo No exitoso N=120**
Media	2.36	2.44
Mediana	1.00	1.00
Moda	1	1
Desviación estándar	1.80	1.71

A las mujeres de ambos grupos les gustaría desarrollar por lo menos una nueva líder en sus grupos de compra al año, y algunas otras opinaron que 2 líderes sería un escenario ideal.

En resumen, los resultados obtenidos en las variables cuantitativas son (Ver Tabla 4.22):

Comparando ambos grupos se observa que en el grupo exitoso su rango de edad va de los 28 a los 42 años, 50 mujeres tienen en **promedio** 4 años de estancia en la compañía y 3 años o menos de ocupar el puesto de líderes.

De estas 50 mujeres la gran mayoría obtuvo el liderazgo al año de haber entrado como distribuidora y sin embargo ya desarrolló en su grupo de compra por lo menos a una líder de ventas, lo cual significa que son mujeres que prometen tener un gran desarrollo en su carrera ejecutiva en la compañía K.

Por el contrario en el grupo no exitoso su rango de edad fluctúa entre los 30 a 54 años, 80 mujeres tienen de permanencia en la compañía un tiempo promedio de 5 años o más y ocupando el puesto de líderes apenas 2 años. Este grupo todavía no desarrolla a ninguna nueva líder de ventas en sus grupos de compra.

Ambos grupos coinciden en que el desarrollar en promedio a una nueva líder en sus grupos al año, sería un excelente resultado para ellas y la compañía en general, ya que el efecto de crecimiento, genera que mayor número de mujeres deseen seguir el ejemplo, así mismo se incrementarían los beneficios (económicos, laborales y morales) para toda la fuerza de ventas.

Tabla 4.22 Resumen de características de ambos grupos

Variables Cuantitativas	**Grupo exitoso**	**Grupo No exitoso**
Tiempo promedio de permanencia en la compañía K	4	5
Tiempo promedio de ser líder de ventas en la compañía K	3	2
Edad promedio	35	41
Promedio actual de líderes desarrolladas en el grupo de compra de la entrevistada	1	0
Promedio de líderes de venta que le gustaría desarrollar en su grupo de compra por año	1	1

Este análisis preliminar de la información describe a las mujeres del grupo exitoso y a las del no exitoso en función de cada una de las variables predictoras. Este análisis servirá de comparativo al aplicar el algoritmo CART.

Las variables como *el tiempo que lleva la líder de ventas en la compañía (T_CIA), el tiempo de ser líder de ventas (T_LIDER), número actual de líderes de venta desarrolladas en el grupo de compra (NOLGPO) y el factor principal que influye para no llegar a ser líder de ventas (FACT_IMP),* son variables que definieron a los grupos (exitoso y no exitoso) y proporcionan el punto de vista de las líderes actuales, por lo que estas variables no se introdujeron en el algoritmo CART.

Aplicando el principio de parsimonia en la construcción del modelo, se modificó la codificación de algunas variables, lo cual no repercute en la generación del modelo, sólo simplifica la categorización de variables y la estandarización de valores. Las variables modificadas son:

Tabla 4.23 Modificación de la codificación de variables

Variable	Codificación anterior	Codificación nueva	Observaciones
FACT_INF	1 = Necesidad económica 2 = Independencia laboral 3 = Confianza en la compañía 4 = Los hijos 5 = Autorrealización 6 = Un cambio de vida 7 = Reconocimiento profesional 8 = Ser una ejecutiva independiente 9 = Timidez 10 = Pareja en desacuerdo 11 = Interés único por las ventas 12 = No había quien cuidara a los hijos 13 = Trabajo adicional absorbente 14 = Desconocimiento de beneficios 15 = Sin experiencia 16 = Falta de decisión 17 = Falta de disciplina 18 = Falta de capacitación	1 = Necesidad económica 1 = Independencia laboral 1 = Confianza en la compañía 1 = Los hijos 1 = Autorrealización 1 = Un cambio de vida 1 = Reconocimiento profesional 1 = Ser una ejecutiva independiente 0 = Timidez 0 = Pareja en desacuerdo 0 = Interés único por las ventas 0 = No había quien cuidara a hijos 0 = Trabajo adicional absorbente 0 = Desconocimiento de beneficios 0 = Sin experiencia 0 = Falta de decisión 0 = Falta de disciplina 0 = Falta de capacitación	De 18 categorías iniciales, ahora se presentan 2, clasificadas como: 1= Factor positivo[32] 0= Factor negativo
TOMAR_DE	1 = Sí, le fue difícil tomar la decisión. 2 = No, le fue difícil tomar la decisión.	1 = Sí, le fue difícil tomar la decisión. 0 = No, le fue difícil tomar la decisión.	El cambio, se realizó para estandarizar a valores de 0,1.
EXP_VENT	1 = Sí, tiene experiencia 2 = No, tiene experiencia	1 = Sí, tiene experiencia 0 = No, tiene experiencia	El cambio, se realizó para estandarizar a valores de 0,1.
EXPCIA	1 = Siempre 2 = Casi siempre	1 = Siempre 0 = Casi siempre	El cambio, se realizó para estandarizar a valores de 0,1.

[32] La clasificación de las categorías de la variable FACT_INF como factor positivo y negativo, fue sugerida por las líderes de venta.

Capítulo 5

CONSTRUCCIÓN DE LOS MODELOS, UTILIZANDO EL ALGORITMO CART

Palabras Clave: Algoritmo CART. The Model Setup. Navigator.

5.1 Especificaciones del CART v 5.0 para la generación de modelos

Para la construcción de modelos, CART presenta el siguiente procedimiento:

1) La definición de los parámetros y variables iniciales que intervendrán para la construcción del árbol, así como su proceso de generación, se lleva a cabo a través de la pantalla “The Model Setup”.
2) La pantalla donde se muestra el árbol y sus reportes correspondientes se denomina “Navigator”.

Cada una de estas pantallas tiene distintas opciones, cada una le da posibilidades al usuario de modificar y experimentar con el modelo tantas veces como sea necesario.

En este capítulo se detallarán las opciones tanto de “The Model Setup” como de “Navigator”, además de presentar la propuesta de configuración inicial de *cinco modelos* para el tema de investigación. La base teórica de la metodología del algoritmo CART, se puede consultar el tercer capítulo, este apartado únicamente muestra el caso práctico.

Antes de iniciar la configuración de parámetros y variables del modelo, hay que definir el área de trabajo de esta investigación (Tabla 5.1), éstas cuestiones básicas las exige la fase de minería de datos (ver capítulo II), para que los modelos generados sean inteligibles.

Tabla 5.1 Definición de parámetros de acuerdo a la fase de minería de datos

Tarea:	***Clasificación.*** *De un grupo de nuevas distribuidoras, el modelo clasificará aquellas distribuidoras que obtendrán el liderazgo de ventas en un año, así como aquellas que no lo hagan asignándoles a cada una la etiqueta o clase que corresponda.*
Modelo:	***Predictivo.*** *Generará un modelo de árbol que clasificará o asignará una clase a grupos de datos futuros, tomando como base el modelo aprendido con el conjunto de datos de prueba.*
Método o técnica:	*Árboles clasificadores. Técnica por excelencia para la tarea de clasificación con atributos categóricos.*
Algoritmo:	*CART*
Software:	*CART V 5.0*
Archivo:	*lideres_modelo_cart.sav*

Las especificaciones anteriores deben estar bien comprendidas para el usuario, ya que son la base de los modelos. Una vez asentado lo anterior, se continuará a detallar las opciones de “The Model Setup” como de “Navigator”.

5.2 The Model Setup

The Model Setup (figura 5.1), presenta once opciones *(Model, Categorical, Testing, Selelct Cases, Best Tree, Combine, Meted, Advanced, Cost, Prior y Penalty)* donde se realiza la configuración de variables y parámetros de cada modelo.

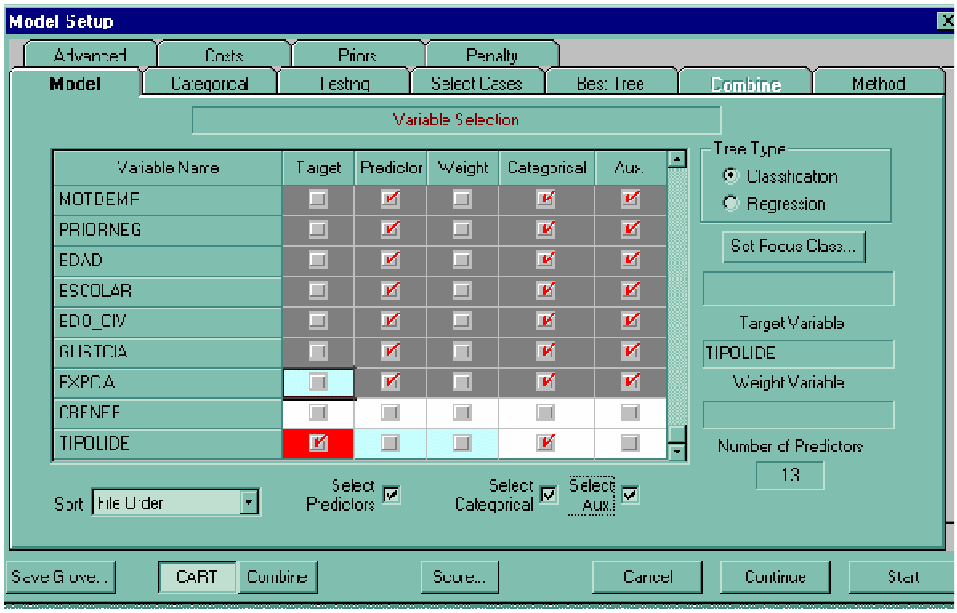


Figura 5.1 *Pantalla The Model Setup*

Descripción de las opciones en “The Model Setup”

1. Model.- Se definen las variables dependientes o de criterio *(target)* e independientes o predictoras *(predictor)*, el tipo de cada una de ellas *(categóricas y/o auxiliares)* y el tipo de árbol de que se construirá *(clasificación o regresión).*

El tipo de la variable dependiente *(target)*, determinará el tipo del árbol. Si la variable es *categórica*, el árbol es de *Clasificación* y si es *continua* será de *Regresión*.

Para generar el árbol de acuerdo a las especificaciones, debe seleccionarse el botón **Start**, localizado en la parte derecha inferior de la pantalla.

Para esta investigación:

- La *variable de criterio o de respuesta* (*target*) es:

Tipo de líder (TIPOLIDE)

Variable con dos categorías:

TIPOLIDE =1, obtiene el liderazgo de ventas en un tiempo $\leq$ 1 año.

TIPOLIDE = 0, obtiene el liderazgo de ventas en un tiempo > 1 año.

- *Las trece variables independientes o predictoras*[33] para el estudio son: *HIJOS, FACT_INF, TOMAR_DE, PERS_INF, EXP_VENT, EMPLEO_2, MOTDEMP, PRIORNEG, EDAD, ESCOLAR, EDO_CIV, GUSTCIA y EXPCIA.*

Cada una de éstas variables son del tipo *categórico y auxiliar.*

El término “variable auxiliar”, significa que CART obtendrá estadísticos básicos, como *las distribuciones de frecuencias* para variables *categóricas* y *medias y varianzas* para variables *continuas* para cada nodo.

2. Categorical.- Se asignan *nombres* correspondientes a cada una de las variables categóricas, lo cual se reflejará en la gráfica del árbol. Este procedimiento es opcional, en caso de no especificar *los nombres*, CART asigna por *default números* a

[33] La definición de las variables se encuentra en el capítulo 4.

cada una de las categorías que conforman a las variables. El asignar *nombres* especiales a las categorías, da como resultado árboles más comprensibles y didácticos.

El procedimiento consiste en la elección de la variable y después se especifican *los nombres* para cada categoría, como se muestra en la figura 5.2.

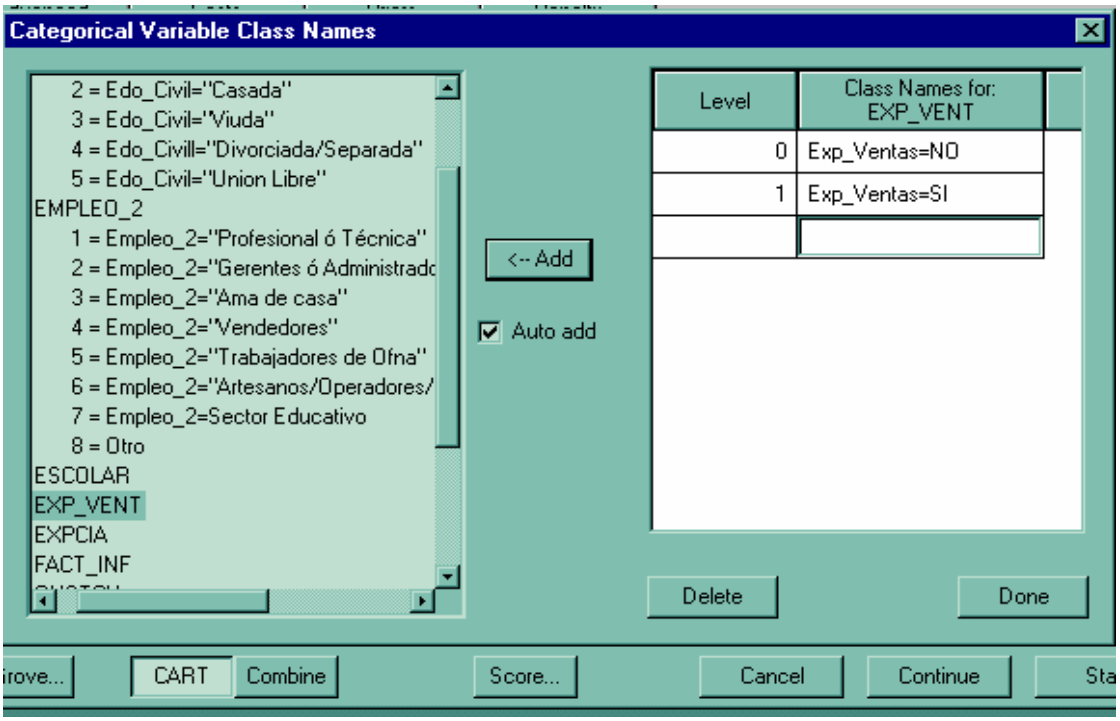


Figura 5.2 Asignación de nombres a categorías

3. Testing. La determinación del *mejor árbol* se hace por las adecuadas estimaciones de los costos que surgen de las clasificaciones erróneas. CART v5.0, incluye 5 métodos de evaluación (*self-validation*) Ver figura 5.3.

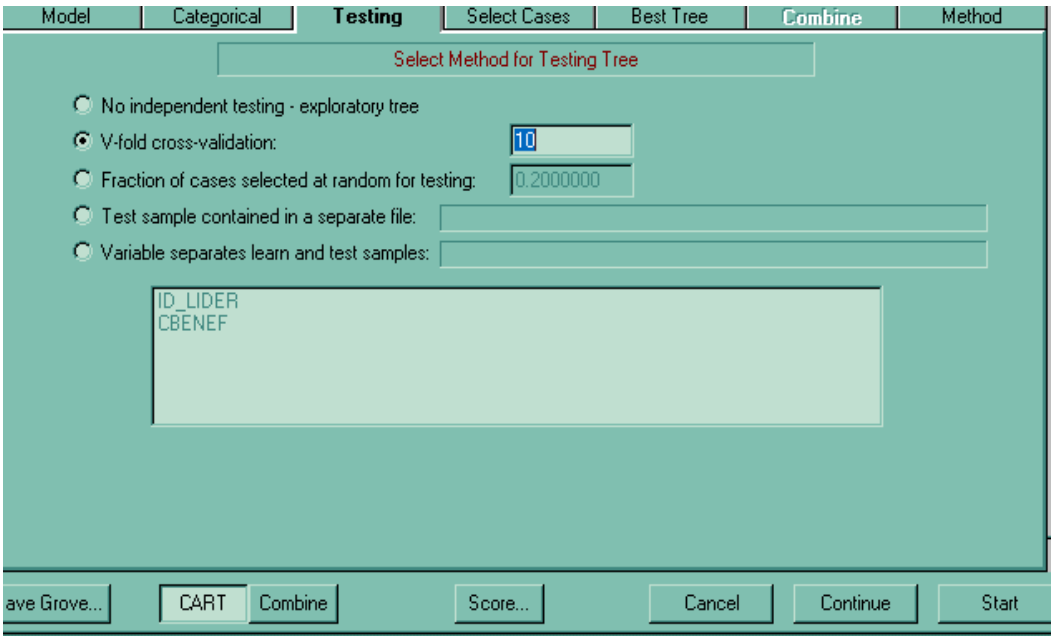


Figura 5.3 Opción Testing

3.1 ***No independent testing.*** Esta opción cancela la fase de prueba y siempre presenta "el árbol más grande", es por ello que su utilización se recomienda en las etapas tempranas del análisis, y le sirve al usuario para familiarizarse con el conjunto de datos. Puede llegar a utilizar muchos recursos de cómputo.

3.2 **V-fold cross validation (Opción por default V=10).** Se utiliza cuando el conjunto de datos no es grande, es decir, el total de registros $\leq$ 3,000. Este método, es parsimonioso con los datos, por ello utiliza con mayor efectividad todos los datos y proporciona mayor estabilidad a la estructura del árbol, computacionalmente es más costoso. Dado que el total de registros empleados en esta investigación es de 200, se empleó esta opción.

El algoritmo crea dos subconjuntos de forma aleatoria, uno es para aprendizaje y el otro es para evaluación.

La muestra de aprendizaje original L, se divide aleatoriamente en V subconjuntos, $L_v, v = 1,...,v$ conteniendo cada uno el mismo número de casos (tanto como sea posible).

La v^{th} muestra de aprendizaje es: $L^{(v)} = L - L_v, v = 1,...,v,$ tal que $L^{(v)}$ contiene la fracción $(V-1)/V$ del total de los datos. Usualmente V es tomado como 10, así que cada muestra de aprendizaje $L^{(v)}$ contiene $9/10$ de los casos.

En *V-fold cross-validation,* V construye árboles auxiliares del árbol principal L, el v^{th} árbol auxiliar crece usando la muestra de aprendizaje $L^{(v)}$.

En los trabajos de investigación, ejecutando casos con $V=10$ se ha obtenido una eficiencia adecuada; aunque en algunos casos se han tomado valores más pequeños de V, también se han obtenido buenos resultados, no así cuando se ha tomado a V más grande de 10.

3.3 Fraction of cases selected at random for testing (default set to 0.20).

Únicamente trabaja con la muestra de *prueba o validación*.

3.4 Test sample conteined in a separate file*. Las muestras de aprendizaje y de prueba son separadas. El conjunto de datos es grande.*

3.5 Variable separates learn and test samples (binary indicator). Utiliza variables binarias de control (1,0) para las muestras de aprendizaje y prueba. Cualquier observación con el valor de la variable igual al valor designado será asignado al conjunto de prueba.

4. Select Cases. Está opción hace que la construcción del árbol se realice sobre la base de un subconjunto de casos. El subconjunto de casos, resulta de seleccionar las variables que se desean que aparezcan en el conjunto de datos.

Por ejemplo, se puede requerir que para la construcción del árbol se necesitan excluir todos aquellos registros donde la edad sea menor o igual a 35 años.

En nuestro caso no se empleó ningún criterio de exclusión.

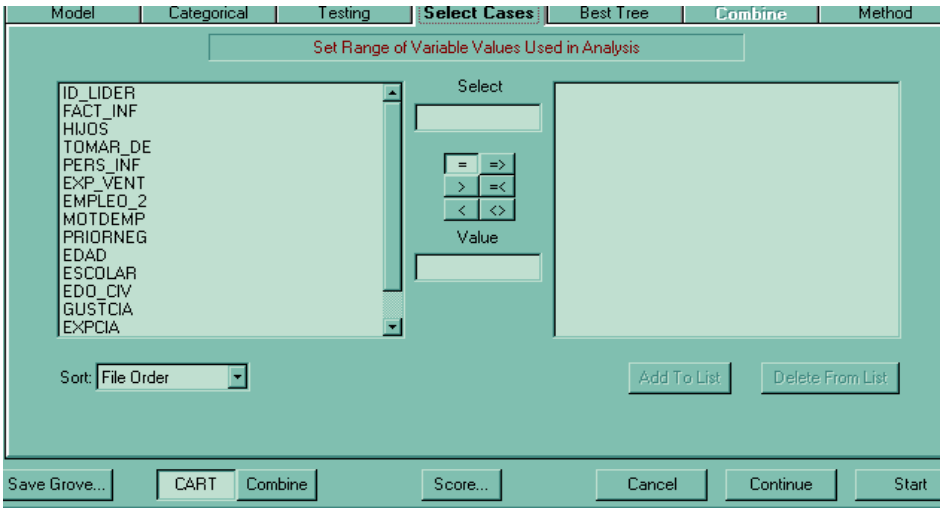


Figura 5.4 Opción Select Cases

5. Best Tree. El algoritmo permite modificar algunos parámetros para influenciar en la construcción de ramas y obtener el mínimo costo para la tasa de clasificaciones erróneas. Se optó por dejar los parámetros de *default*:

- *Regla del Error Estándar (Standard Error Rule),* aquí se especifica el parámetro que el algoritmo utilizará para seleccionar “el mejor árbol”. El parámetro que por *default* se utiliza es *el costo mínimo* a pesar del tamaño del árbol *(the Minimun cost tree regardless of size).*

- *Variable de Importancia en la fórmula (Variable importance formula*), se asignan puntajes a las variables de importancia, con la finalidad de mejorara el método de construcción de ramas y afinar el método.

- *All surrogates count equally*, se refiere a los valores que se asignarán a ramas sustitutas en cada nodo.

6. Method. Especificación de las *reglas de ramificación,* estas reglas son el ingrediente fundamental del crecimiento del árbol. Se tienen seis índices de diversidad o medidas de impureza diferentes para la *clasificación*, cada uno con sus propias *reglas: Gini, Symmetric Gini, Entropy, Class Probability, Twoing y Ordered Twoing.*

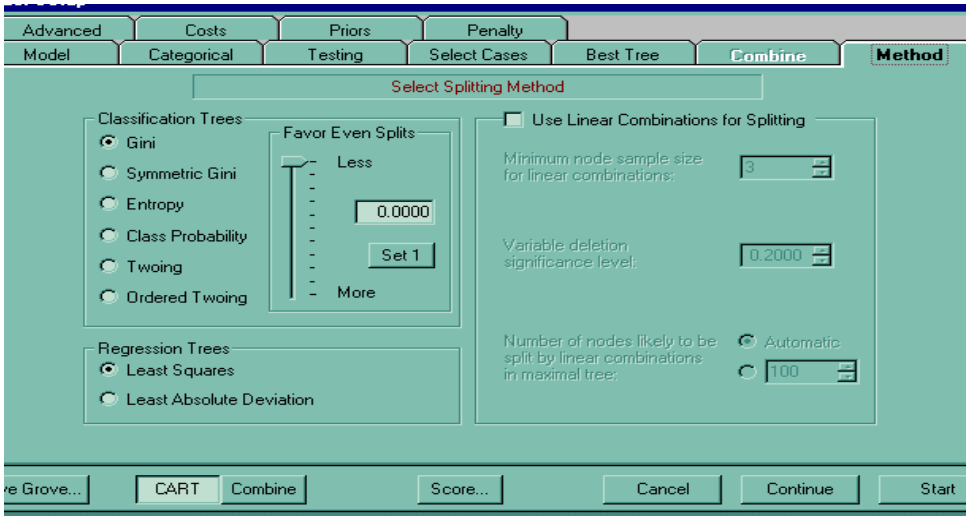


Figura 5.5 Opción Method

7. Cost. Establecimiento de *los parámetros de costos y penalizaciones* en la construcción del árbol. Para *árboles de clasificación*, se especifican costos para clasificaciones erróneas, es decir, ¿Cuánto cuesta clasificar erróneamente un registro?, los valores por *default* para todos los costos es de uno, esto se representa en una matriz de costos, como se ilustra en la figura 5.6.

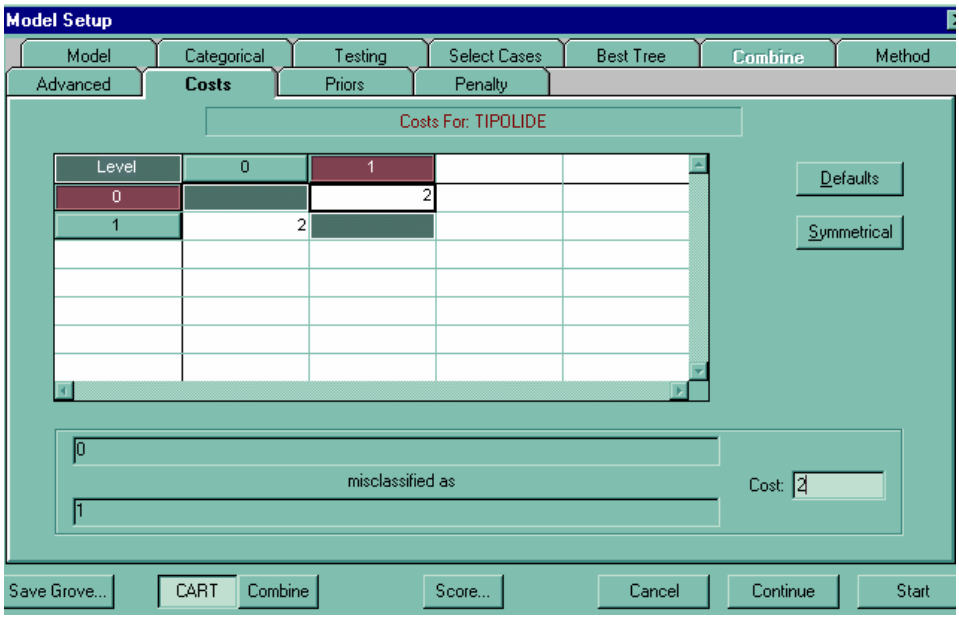


Figura 5.6 Opción Cost

En el caso de la investigación, *¿cuánto cuesta clasificar a la variable TIPOLIDE=1, cuando debió clasificarse como TIPOLIDE=0?, es decir clasificar a una distribuidora como una mujer que obtendrá el liderazgo de ventas en un tiempo menor o igual a un año, debiendo ser lo contrario.* En la investigación se optó por dejar los costos por *default.*

8. Priors. En la opción de *Prioridades*, se le da forma al análisis de *clasificación,* es decir, CART, puede asumir que las clases sean tratadas como si estuvieran uniformemente distribuidas en la población a pesar de su distribución real en la muestra. Los parámetros por *default*, frecuentemente dan resultados satisfactorios porque cada clase es tratada con igual importancia para clasificaciones veraces.

De las 6 diferentes alternativas que ofrece la herramienta, se eligió el parámetro por *default* (EQUAL), las otras son: *DATA, MIX, LEARN, TEST y SPECIFY*.

9. Penalti. Se especifican los *parámetros de penalizaciones* en las variables con valores erróneos o faltantes, puede establecerse que *no haya penalidad o que exista una alta penalidad.* Las *penalizaciones* se asignan para valores incorrectos *(variables independientes categóricas o continuas)* y para cuando hay un alto número de niveles *(solo variables categóricas)*.

Penalidad para las variables: cada penalidad se transforma en un factor multiplicativo entre 0 y 1 aplicado para mejorar las puntuaciones de las variables *predictoras o independientes* en la construcción del árbol.

5.3 Navigator

La pantalla “Navigator”, muestra el árbol generado y los reportes correspondientes al modelo.

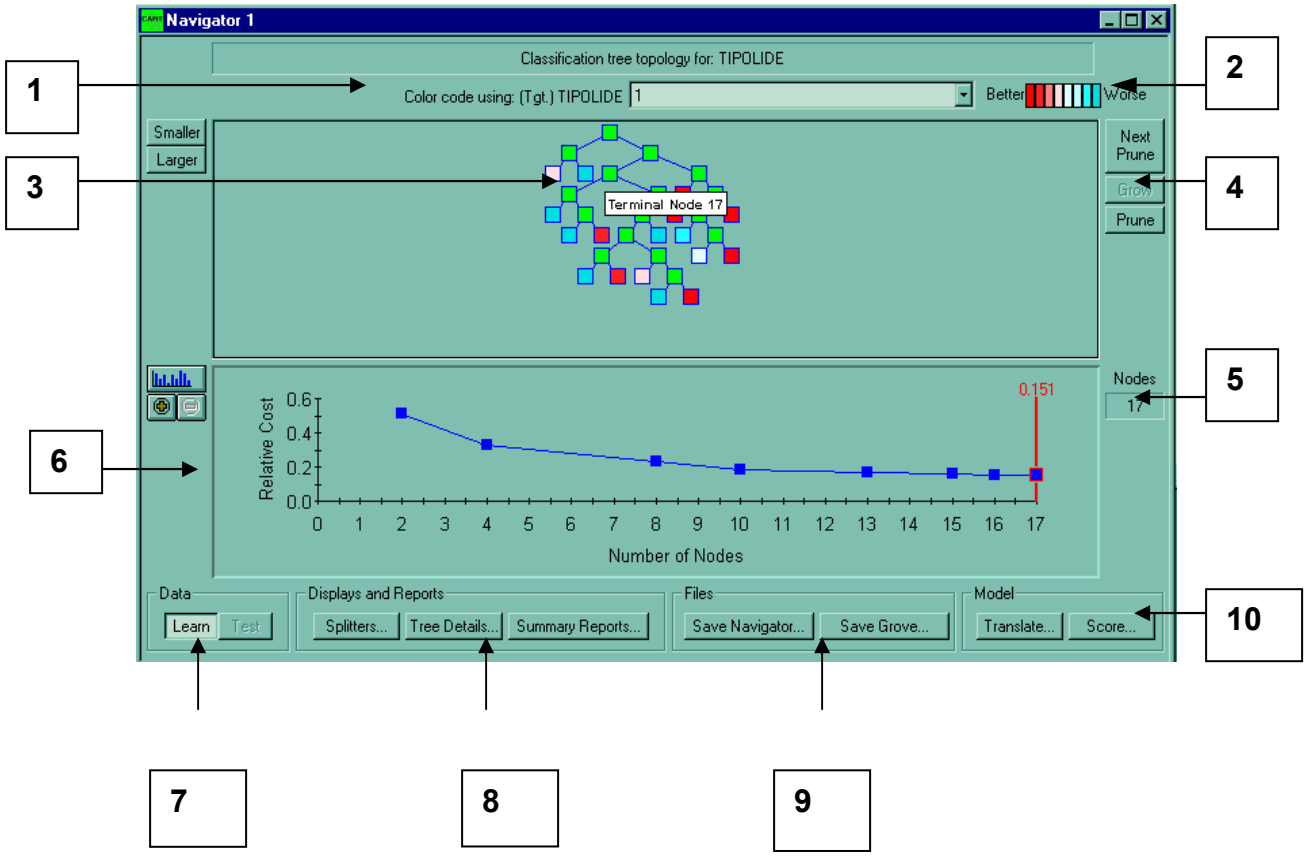


Figura 5.8 La pantalla Navigator muestra la salida del árbol

Descripción de las opciones en “Navigator”

Cada una de las opciones se encuentra ilustrada con números en la figura 5.8, la explicación de ellas se presenta a continuación:

1) Se muestran las clases que corresponden a la variable de criterio o dependiente (*target*). En nuestro caso la variable *TIPOLIDE*, puede tomar 2 valores [1,0]. Para cada una se generan árboles, muy parecidos. En este estudio se trabaja cuando la variable *TIPOLIDE=1 (clase1).*

2) Muestra el código de color para los nodos terminales. Los nodos terminales están codificados con un color, el cual indica si una clase en particular mejora (rojo) o empeora (azul) con respecto al nodo terminal más puro cuando se compara con el nodo raíz.

3) La gráfica del árbol se despliega en su totalidad y, cada nodo tiene información acerca de las clases.

4) Opciones para realizar el procedimiento de poda.

5) Indica el número total de nodos de los que se compone el árbol, este varia según el procedimiento realizado (podas).

6) Se muestran las gráficas de costos y ganancias (gains), y de porcentajes de población en cada nodo (terminal nodes).

7) De acuerdo al método de construcción del árbol seleccionado, los datos con los que se puede estar trabajando definirá el procedimiento si es de *aprendizaje o de validación.*

8) Los reportes detallados consisten en:

a) El diagrama de ganancia, muestra la contribución de los nodos para cubrir una clase en particular. Los nodos se ordenan con base al porcentaje de casos de la clase uno y va del más alto al más bajo.

b) La distribución de los nodos terminales, es un diagrama de barras y cada barra representa un nodo terminal, el orden que sigue va del nodo *más rico o alto en población en la clase al más bajo.*

c) Variables de Importancia, se presentan las variables junto con su puntaje, el cual refleja la contribución que cada variable hace en la clasificación o en la predicción de la variable dependiente.

d) Tablas de clasificaciones erróneas, aquí se muestra como muchos casos fueron clasificados incorrectamente en el árbol, tanto para los ejemplos de aprendizaje como de validación.

e) Tablas de predicción o matriz de confusión, muestra si CART tiende a concentrar las clasificaciones erróneas en clases específicas y si así fuera donde ocurrirían.

9) Opciones para el guardado de los archivos.

10) Opciones de los archivos generados con los modelos finales.

Capítulo 6

EVALUACIÓN DE LOS MODELOS Y ANÁLISIS DE RESULTADOS

Palabras Clave: Muestra de aprendizaje. Muestra de validación. Reglas de Clasificación. Patrones de comportamiento.

En este capítulo se evaluarán cinco modelos, con la finalidad de observar el comportamiento del algoritmo en cada uno y elegir el que mejor describa el comportamiento de los datos, es decir, *el modelo que aprendió de una muestra de aprendizaje el comportamiento y lo aplicará a grupos de datos futuros de nuevas distribuidoras con la mayor precisión posible.*

6.1 Propuesta de modelos

Los parámetros de configuración de los cinco modelos se muestran en la Tabla 6.1.

Tabla 6.1 Modelos propuestos

Parámetros de configuración	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
1) Nivel de la variable dependiente evaluada	Tipolide = 1	Tipolide = 1	Tipolide = 1	Tipolide = 1	Tipolide = 1
2) No. de variables independientes	13	13	13	13	13
3) Evaluación de las muestras de aprendizaje y validación *(Testing)*	*No independent testing – exploratory tree*	*V-fold cross-validation (V-fold=10)*	*V-fold cross-validation (V-fold=10)*	*V-fold cross-validation (V-fold=10)*	*V-fold cross-validation (V-fold=10)*
4) Selección de casos (Select Cases)	Ninguna	Ninguna	Ninguna	Ninguna	Ninguna
5) Mejor árbol (Best Tree)	Default	Default	Default	Default	Default
6) Método de generación de reglas (Method)	GINI	GINI	GINI	GINI	GINI
7) Opciones avanzadas (Advanced)	Default	Default	Default	Default	Default
8) Costos (Cost)	*Default* (Cost=1.0)	*Default* (Cost=1.0)	*Default* (Cost=1.0)	*Default* (Cost=1.0)	*Default* (Cost=1.0)
9) Prioridades (Priors)	EQUAL (default)	EQUAL (default)	EQUAL (default)	EQUAL (default)	EQUAL (default)
10) Penalidades (Penalty)	Default (Cero)	Default (Cero)	Default (Cero)	1	1
11) Estrategia de poda	Ninguna	Ninguna	Una	Ninguna	Una

El primer modelo, representa el árbol completo, donde el algoritmo cancela la fase de prueba con los datos, este modelo sirve como un análisis preliminar.

El segundo modelo utiliza el método de evaluación para muestras de aprendizaje y de validación el V-fold cross-validation (V-fold=10).

El tercer modelo, resulta de aplicar la estrategia de poda al segundo modelo.

Al cuarto modelo además de trabajar con V-fold cross-validation (V-fold=10), se le aplica una penalidad de uno, con el objetivo de mejorar resultados.

El quinto modelo, resulta de aplicar la estrategia de poda al cuarto modelo.

Cada modelo se ejecutó por separado en “The Model Setup”, y cada resultado es independiente.

6.2 Evaluación del árbol completo para análisis preliminar

CART, presenta el árbol donde los nodos terminales se identifican con dos colores diferentes, esto representa el número de *ejemplos* o casos que caen en los nodos y son etiquetados con una clase o etiqueta.

La clase uno identifica *aquellas distribuidoras que obtienen el liderazgo de ventas en un tiempo menor o igual a un año,* y la clase cero las que lo obtuvieron en un tiempo mayor.

Todos modelos se generaron con la variable TIPOLIDE=1. Las demás variables son predictoras o independientes (ver capítulo 5).

En la figura 6.1, se muestra el árbol de clasificación completo que genera el algoritmo, el cual tiene diecisiete nodos terminales.

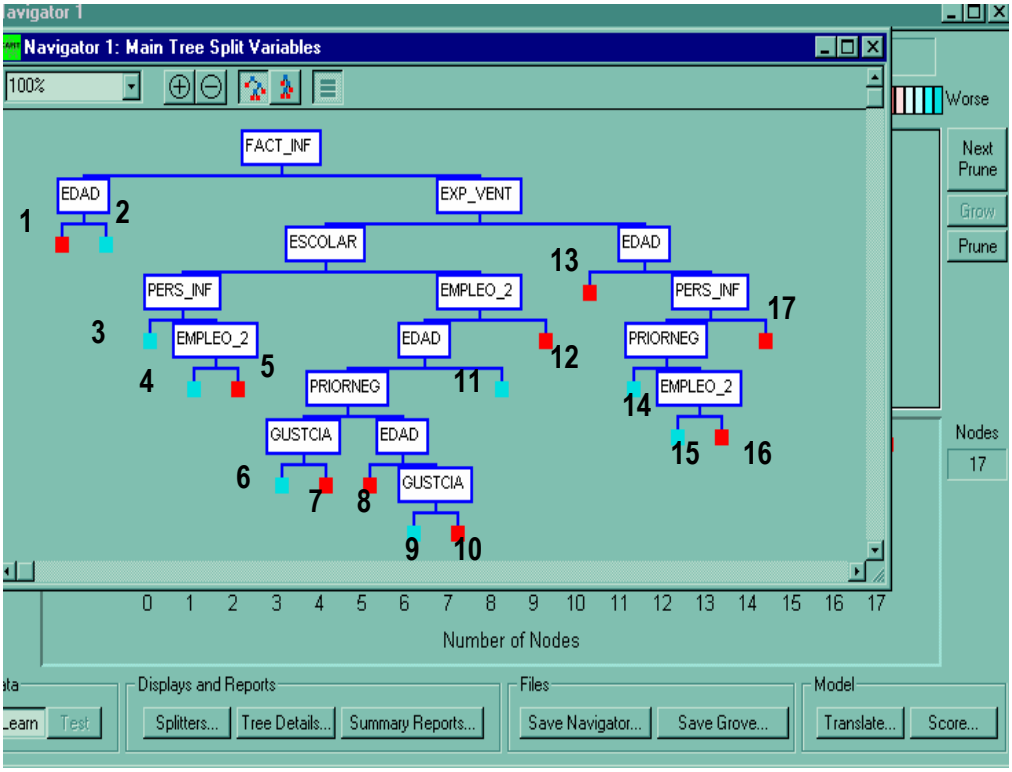


Figura 6.1 *Árbol de clasificación completo para la variable de criterio TIPOLIDE=1*

Se observa en la Tabla 6.2 que el CART identifica al nodo terminal diecisiete como “el mejor”, ya que es un nodo con alta concentración de casos de la clase uno y además obtuvo el menor costo, a pesar de que el nodo trece también es terminal y gana en porcentaje de concentración de casos.

Tabla 6.2 Resumen de nodos terminales

Color del nodo	Nodo terminal	Casos por nodo	% de contribución en el nodo	Pureza del nodo
rojo	**13**	**25**	**100**	Nodos con mayor pureza o con el mayor número de casos que pertenecen a la clase uno
	17	**14**	**93.33**	
	10	11	91.66	
	12	10	90.90	
	16	8	88.88	
	7	3	75	
	5	5	71.42	
rosa	1	1	50	Nodos con pobre concentración de casos de la clase uno
	8	3	42.85	
azul	15	2	40	Casi ningún caso de la clase uno
	14	1	20	
	3	0	4	
	2	0	0	
	6	0	0	
	11	0	0	
	4	0	0	
	9	0	0	

Importancia de las variables en la obtención del liderazgo en un año

La tabla 6.3 proporciona un ejemplo de la importancia de cada variable predictora en la construcción del árbol. El índice de diversidad Gini asigna una puntuación a cada variable para hacer las divisiones del árbol. Esta tabla muestra la primera evaluación para particionar el árbol, éste es un proceso recursivo que se realiza hasta finalizar la construcción.

Tabla 6.3 Importancia de cada variable predictora

Variable	Puntuación	Descripción
FACT_INF	100.00	Factores positivos o negativos que influyeron en la toma de decisión
PRIORNEG	80.04	Prioridad del negocio para la líder
EDO_CIV	79.36	Estado civil de la líder cuando alcanzó este nivel
EDAD	60.90	Edad de la líder cuando alcanzó este nivel
EMPLEO_2	50.31	Empleo adicional de la líder
ESCOLAR	39.07	Grado máximo de escolaridad de la líder cuando alcanzó este nivel
EXP_VENT	38.25	Evaluación de la experiencia en ventas de la líder
GUSTCIA	35.78	Factor de la compañía K de mayor agrado para la líder
PERS_INF	28.62	Persona que influyó determinantemente en la toma de la decisión
TOMAR_DE	15.44	Evaluación de la toma de decisión para ser líder en un año
MOTDEMP	12.26	Motivo principal para dejar el empleo adicional
EXPCIA	3.26	Evaluación de las expectativas de la líder sobre la empresa K
HIJOS	2.81	Evaluación de los hijos que son dependientes económicos de la líder

Según se observa, la variable FACT_INF (factores positivos o negativos que influyeron en la toma de decisión), fue la que obtuvo *mayor porcentaje en la evaluación del índice Gini* y por tanto fue elegida para hacer la primera partición del árbol.

Las variables con menor porcentaje en la evaluación fueron EXPCIA (evaluación de las expectativas de la líder sobre la empresa K) e HIJOS (evaluación de los hijos que son dependientes económicos de la líder).

Resultados

Reglas de clasificación para las distribuidoras que obtienen el liderazgo de ventas en un año

Las principales características o el patrón que siguen estas distribuidoras es:

Nodo 13: La distribuidora presenta alguna o varias de las siguientes características: necesidad económica, independencia laboral, confianza en la compañía, los hijos, autorrealización, tener un cambio de vida, reconocimiento profesional, ser ejecutiva independiente; además la mujer tenía experiencia en ventas y su edad promedio se encuentra entre los 29 y 41 años ó

Nodo 17: además de lo mencionado en el nodo anterior, se adiciona que la persona que más influyó en la toma de decisión fue su líder directa y sus hijos.

Reglas de clasificación para las distribuidoras que no obtienen el liderazgo de ventas en un año

Nodo 1 y 2: La distribuidora presenta alguna o varias de las siguientes características: timidez, pareja en desacuerdo, interés único por las ventas, trabajo alterno absorbente, desconocimiento de los beneficios, sin experiencia en ventas, falta

de decisión, falta de disciplina, falta de capacitación, no tener quien cuide a sus hijos; además que su edad sea menor a 20 años o superior a los 40 años.

Validación de resultados

El método *Testing* (evaluación de las muestras de aprendizaje y validación) utilizado en este primer modelo fue el *No independing testing – exploratoria tree*, el cual únicamente evalúa la muestra de aprendizaje. Para evaluar la capacidad del algoritmo CART al momento de predecir la obtención o no del liderazgo de ventas, se puede observar en la Tabla 6.4 lo siguiente:

Tabla 6.4 Clasificaciones correctas e incorrectas en el primer modelo

Clases	Casos	Casos clasificados correctamente	% de clasificación correcta	Casos clasificados incorrectamente	% de clasificación incorrecta	Casos predichos
1	80	80	100	4	5	92
0	120	104	86.66	12	10	108
Total	200	184	92	16	8	200

La proporción de casos mal clasificados fue 4 de 80 que obtuvieron el liderazgo en un tiempo menor o igual a un año (clase uno), representando el 5% y 12 de 120 casos que no obtuvieron el liderazgo, representado el 10 % (clase cero). Esto da como resultado un porcentaje de clasificación correcta del 92% del total de distribuidoras que entraron a la compañía, y lograron el liderazgo de ventas.

Conclusión: El modelo obtenido por el algoritmo CART tiene alto poder predictivo, o que el árbol de clasificación se ajustó totalmente al conjunto de datos y no obtuvo un modelo predictivo, por esta situación se evaluarán otros cuatro modelos para

elegir aquel que presente *el mejor patrón de comportamiento de las distribuidoras ante la opción de elegir el liderazgo de ventas o no*.

6.3 Evaluación del segundo árbol

En la figura 6.2 se presenta el árbol de clasificación con ocho nodos terminales.

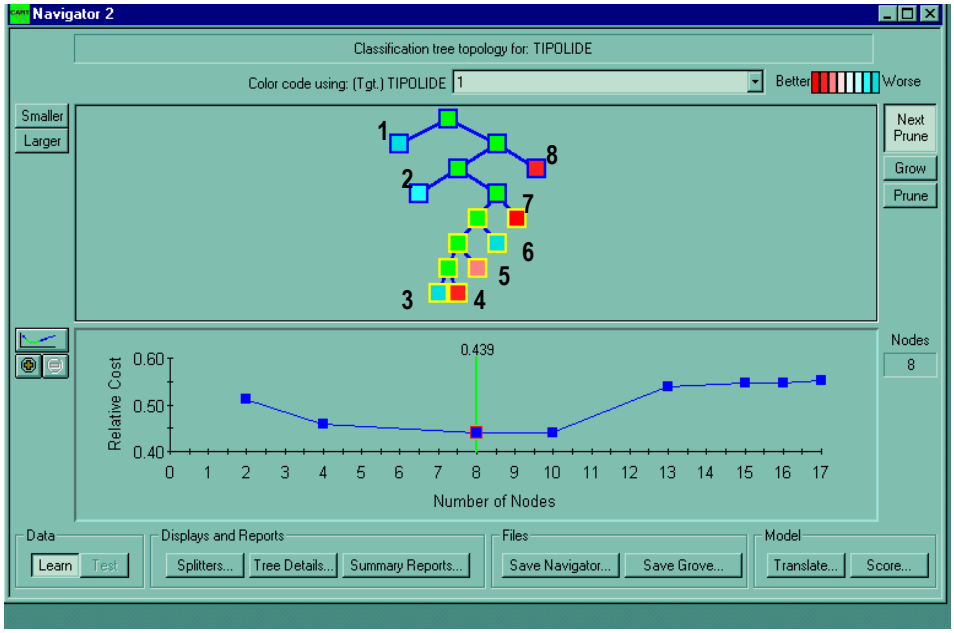


Figura 6.2 *Segundo árbol de clasificación*

CART identifica al nodo terminal al número ocho como “el mejor”, ya que en él se concentra el mayor número de casos de la clase uno.

La tabla 6.5 presenta un resumen del número de casos en los nodos terminales.

Tabla 6.5 Resumen de nodos terminales

Color del nodo	Nodo terminal	Casos por nodo	% de contribución en el nodo	Pureza del nodo
rojo	**7**	**10**	**90.909**	Nodos con mayor pureza o con el mayor número de casos que pertenecen a la clase uno
	8	**50**	**84.746**	
	4	3	75.000	
rosa	5	14	70.000	Nodos con pobre concentración de casos de la clase uno
azul	2	6	16.667	Casi ningún caso de la clase uno
	1	1	1.695	
	3	0	0.000	
	6	0	0.000	

Importancia de las variables en la obtención del liderazgo en un año

La tabla 6.6 proporciona la importancia que tiene cada variable predictora en la construcción del árbol. El índice Gini asignó la puntuación a cada variable para hacer las divisiones correspondientes en el árbol.

Tabla 6.6 Importancia de cada variable predictora

Variable	Puntuación	Descripción
FACT_INF	100.00	Factores positivos o negativos que influyeron en la toma de decisión
EDO_CIV	73.15	Estado civil de la líder cuando alcanzó este nivel
PRIORNEG	63.07	Prioridad del negocio para la líder
EXP_VENT	38.25	Evaluación de la experiencia en ventas de la líder
EDAD	38.16	Edad de la líder cuando alcanzó este nivel
ESCOLAR	27.80	Grado máximo de escolaridad de la líder cuando alcanzó este nivel
EMPLEO_2	27.61	Empleo adicional de la líder
GUSTCIA	18.43	Factor de la compañía K de mayor agrado para la líder
TOMAR_DE	15.44	Evaluación de la toma de decisión para ser líder en un año
PERS_INF	11.13	Persona que influyó determinantemente en la toma de decisión
MOTDEMP	6.37	Motivo principal para dejar el empleo adicional
HIJOS	2.81	Evaluación de los hijos que son dependientes económicos de la líder
EXPCIA	0.00	Evaluación de las expectativas de la líder sobre la empresa K

Nuevamente la variable FACT_INF (factores positivos o negativos que influyeron en la toma de decisión), fue la que obtuvo mayor porcentaje en la evaluación del índice Gini y la variable con cero porcentaje fue EXPCIA (evaluación de las expectativas de la líder sobre la empresa K).

Resultados

Reglas de clasificación para las distribuidoras que obtienen el liderazgo de ventas en un año

El principal patrón que siguen estas distribuidoras es:

Nodo 8: Las distribuidoras presentan alguna de estas características: necesidad económica, independencia laboral, confianza en la compañía, los hijos, autorrealización, tener un cambio de vida, reconocimiento profesional, ser ejecutiva independiente; además la mujer tenía experiencia en ventas ó

Nodo 7: además de lo mencionado en el nodo anterior, se adiciona que la mujer tenga uno de los siguientes grados de escolaridad: secundaria, técnico o comercio, bachillerato, licenciatura o postgrado y además que su actividad laboral fuera: ama de casa, trabajadora de oficina, profesional o técnica, sector educativo, gerentes o administradoras.

Reglas de clasificación para las distribuidoras que no obtienen el liderazgo de ventas en un año

Nodo 1: La distribuidora presenta alguna o varias de las siguientes características: timidez, pareja en desacuerdo, interés único por las ventas, trabajo alterno absorbente, desconocimiento de los beneficios, sin experiencia en ventas, falta de decisión, falta de disciplina, falta de capacitación y no tener quien cuide a sus hijos.

Nodo 2: Las distribuidoras presentan alguna de estas características: necesidad económica, independencia laboral, confianza en la compañía, los hijos, autorrealización, tener un cambio de vida, reconocimiento profesional o ser ejecutiva independiente. Además de lo anterior la mujer no tiene experiencia en ventas y su instrucción escolar es baja o media como la primaria, la secundaria o inclusive el

bachillerato, pero si la mujer cuenta con un alto grado escolar como un postgrado también es probable que no obtenga el liderazgo.

Validación de resultados

En este modelo, el algoritmo divide el conjunto de datos de forma aleatoria en dos conjuntos de datos, *el de aprendizaje*, con la finalidad de obtener el patrón de comportamiento y aprenderlo para datos futuros y *el de validación*, en el que se evalúa el patrón obtenido y observar el comportamiento del modelo.

Evaluando la capacidad del algoritmo CART en la muestra de aprendizaje, se observa en la Tabla 6.7 que la proporción de casos clasificados erróneamente fue 7 de 80 de la clase uno representando el 8.75% y 17 de 120 de la clase cero representando un 14.16%. Esto da como resultado un porcentaje de clasificación correcta del 88% del total de las distribuidoras que se iniciaron en este negocio y lograron el liderazgo de ventas.

Tabla 6.7 Clasificaciones correctas e incorrectas en la muestra de aprendizaje del segundo modelo

Muestra de aprendizaje						
Clase	**Casos**	**Casos clasificados correctamente**	**% de clasificación correcta**	**Casos clasificados incorrectamente**	**% de clasificación incorrecta**	**Casos predichos**
1	80	73	91.25	7	8.75	94
0	120	103	85.83	17	14.16	106
Total	200	176	88	24	12	200

Evaluando la muestra de validación, se observa en la Tabla 6.8 que la proporción de casos clasificados erróneamente fue 13 de 80 de la clase uno representando el 16.25 % y 33 de 120 de la clase 0 representando un 27.5%. Esto da como resultado un porcentaje de clasificación correcta del 77% del total de las distribuidoras que se iniciaron en este negocio y lograron el liderazgo de ventas.

Conclusión: Comparado con la muestra de aprendizaje el porcentaje de clasificación correcta es más bajo, lo que significa que el modelo aprendido se comporta de forma general y no de forma específica con los datos de prueba, para verificarlo se realizará el procedimiento de poda en este modelo y comparar resultados.

Tabla 6.8 Clasificaciones correctas e incorrectas en la muestra de validación del segundo modelo

Muestra de validación						
Clase	**Casos**	**Casos clasificados correctamente**	**% de clasificación correcta**	**Casos clasificados incorrectamente**	**% de clasificación incorrecta**	**Casos predichos**
1	80	67	83.75	13	16.25	104
0	120	87	72.5	33	27.5	96
Total	200	154	77	46	23	200

6.4 Evaluación del tercer modelo: estrategia de poda aplicada al segundo modelo

Al realizar la estrategia de poda al segundo modelo, las ramas eliminadas son EMPLEO_2 (Empleo adicional de la líder), EDAD (Edad de la líder), PRIORNEG (Prioridad del negocio para la líder), y GUSTCIA (Factor de la compañía K de mayor agrado para la líder) y los nodos terminales eliminados son tres, cuatro, cinco, seis y siete; el nodo ocho, que es considerado inicialmente por el algoritmo como “el mejor” queda intacto. Las ramas conservadas son FACT_INF (Factores Positivos o negativos que influyeron en la toma de decisión), EXP_VENT (Experiencia en ventas) y ESCOLAR (Grado máximo de escolaridad). (Ver figuras 6.3 (a) y 6.3 (b)).

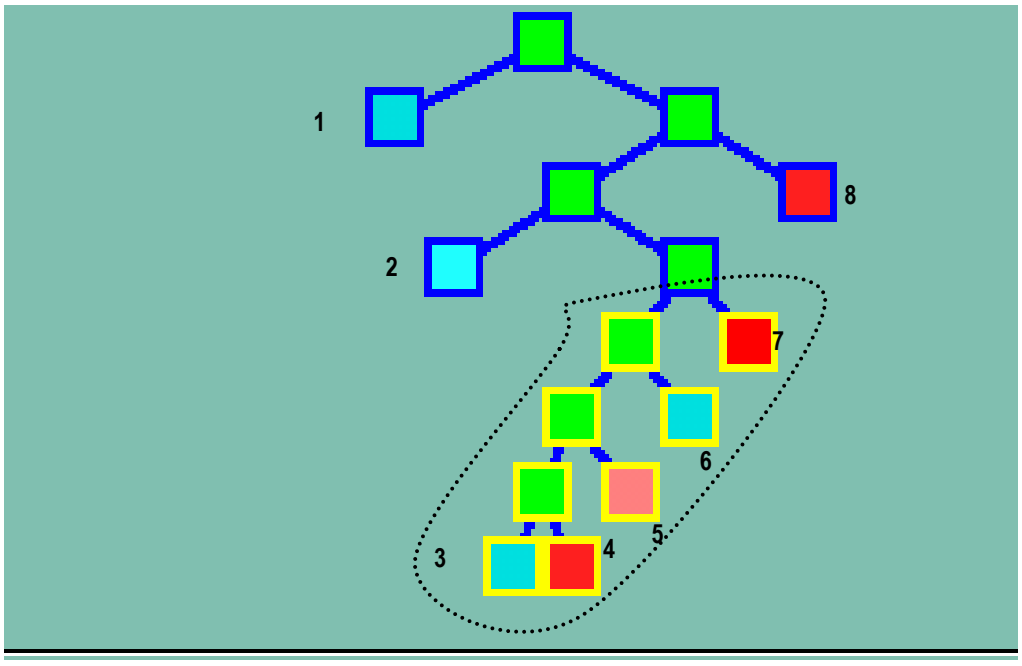


Figura 6.3 (a) *Estrategia de poda en el segundo modelo*

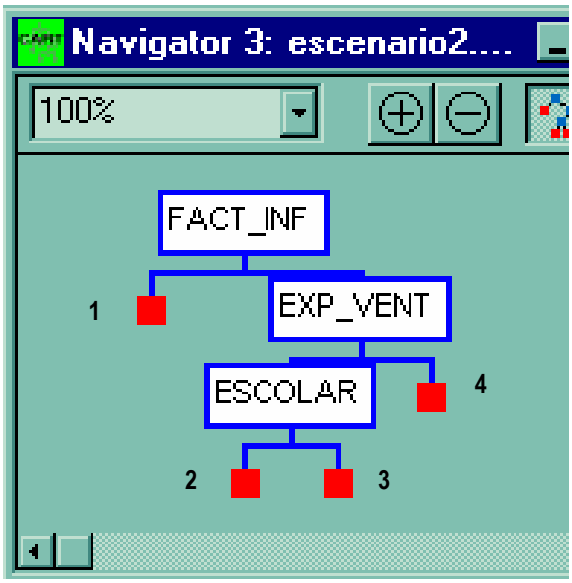


Figura 6.3 (b) *Modelo resultante de la primera poda*

Comparación de nodos terminales con mayor contribución para la clase uno en ambos modelos

En la Tabla 6.9 se puede ver que, tanto en el modelo original como en el de poda, se conservó el porcentaje de casos de la clase uno en el nodo considerado como "el mejor" (ahora nodo cuatro).

Tabla 6.9 Comparación de nodos terminales en ambos modelos

Modelo original		**Modelo con una poda**	
Nodo	**Contribución**	**Nodo**	**Contribución**
7	90.90%	**4**	**84.746%**
8	**84.74%**	3	58.696%
4	75%	2	16.667%
5	70%	1	1.695%
2	16.66%		
1	1.69%		
3	0%		
6	0%		

Comparación de las variables con mayor puntaje para la ramificación

En la Tabla 6.10 se observa que tanto en el modelo original como en el de poda, el algoritmo tomó la mayoría de las variables, excluyendo en el original a EXPCIA y dos en el de poda HIJOS y EXPCIA.

Tabla 6.10 Comparación de las variables predictoras en ambos modelos

	Modelo original		**Modelo con una poda**	
	Variable	**Puntuación**	**Variable**	**Puntuación**
1	FACT_INF	100.00	FACT_INF	100.00
2	EDO_CIV	73.15	EDO_CIV	63.10
3	PRIORNEG	63.07	PRIORNEG	56.53
4	EXP_VENT	38.25	EXP_VENT	38.25
5	EDAD	38.16	ESCOLAR	24.99
6	ESCOLAR	27.80	EDAD	24.79
7	EMPLEO_2	27.61	EMPLEO_2	18.77
8	GUSTCIA	18.43	TOMAR_DE	10.94
9	TOMAR_DE	15.44	PERS_INF	7.44
10	PERS_INF	11.13	GUSTCIA	5.37
11	MOTDEMP	6.37	MOTDEMP	2.16
12	HIJOS	2.81	HIJOS	0.00
13	EXPCIA	0.00	EXPCIA	0.00

Resultados

Las reglas de clasificación tanto para las distribuidoras que obtienen el liderazgo de ventas en un año como para las que no, son las mismas que en el modelo original (antes de aplicar la estrategia de poda).

Validación de resultados

Muestra de aprendizaje

Las tablas 6.11 y 6.12 muestran un comparativo de las clasificaciones erróneas en la *muestra de aprendizaje* tanto en el modelo original como en el modelo con poda.

Tabla 6.11 Comparación de las clasificaciones correctas e incorrectas del segundo modelo y del modelo con poda en la muestra de aprendizaje

Muestra de Aprendizaje						
	Modelo original			Modelo con poda		
Clase	**Casos clasificados correctamente**	**Casos clasificados incorrectamente**	**Casos predichos**	**Casos clasificados correctamente**	**Casos clasificados incorrectamente**	**Casos predichos**
Clase 1 N=80	73	7	94	73	7	105
Clase 0 N=120	103	17	106	92	28	95
Total	176	24	200	165	35	200

Tabla 6.12 Comparación de los porcentajes de las clasificaciones correctas e incorrectas del segundo modelo y del modelo con poda en la muestra de aprendizaje

Muestra de Aprendizaje				
	Modelo original		Modelo con poda	
Clase	**% de clasificación correcta**	**% de clasificación incorrecta**	**% de clasificación correcta**	**% de clasificación incorrecta**
Clase 1 N=80	91.25	8.75	91.25	8.75
Clase 0 N=120	85.83	14.16	76.66	23.33
Total	88	12	82.50	17.50

En el modelo con poda, la proporción de casos mal clasificados en la clase uno, fue 7 de 80, representando el 8.75% y en la clase cero fue 28 de 120, representando el 23.33%, comparado con el modelo original la clase uno no varia su porcentaje de clasificación incorrecta y en la clase cero varía por 9.17%, siendo más alto el porcentaje de error en el modelo con poda. El porcentaje de clasificación correcta en el modelo original es de 88% mientras que en el de poda es de 82.50%.

Muestra de validación

Las tablas 6.13 y 6.14 muestran un comparativo de las clasificaciones erróneas en la *muestra de validación* tanto en el modelo original como en el modelo con poda.

Tabla 6.13 Comparación de las clasificaciones correctas e incorrectas del segundo modelo y del modelo con poda en la muestra de validación.

Muestra de Validación						
	Modelo Original			**Modelo con Poda**		
Clase	**Casos clasificados correctamente**	**Casos clasificados incorrectamente**	**Casos predichos**	**Casos clasificados correctamente**	**Casos clasificados incorrectamente**	**Casos predichos**
Clase 1 N=80	67	13	104	66	14	104
Clase 0 N=120	87	33	96	86	34	96
Total	154	46	200	152	48	200

Tabla 6.14 Comparación de los porcentajes de las clasificaciones correctas e incorrectas del segundo modelo y del modelo con poda en la muestra de validación

	Muestra de Validación			
	Modelo original		**Modelo con poda**	
Clase	**% de clasificación correcta**	**% de clasificación incorrecta**	**% de clasificación correcta**	**% de clasificación incorrecta**
Clase 1 N=80	83.75	16.25	82.50	17.5
Clase 0 N=120	72.50	27.50	71.66	28.33
Total	77	23	76	24

En el modelo con poda, la proporción de casos mal clasificados en la clase uno, fue 14 de 80, representando el 17.5% y en la clase cero fue 34 de 120, representando el 28.33%, comparado con el modelo original la clase uno varia su porcentaje de clasificación incorrecta en 1.25% y en la clase cero varía por 0.83%, siendo más alto el porcentaje de error en el modelo con poda. El total de clasificaciones correctas en el modelo original es del 77% y en el de poda es del 76%

Conclusión: Se observa que el comportamiento del modelo tanto en el original como en el de poda es muy semejante en ambas muestras (aprendizaje y validación), esto significa que en este caso la estrategia de poda no afectó el comportamiento del modelo.

Se evaluará un tercer modelo *asignando al parámetro de penalidad el valor de uno,* esto significa que se obliga al algoritmo a ser más asertivo en el criterio de selección de ramas y lo obliga a cometer menos clasificaciones erróneas.

6.5 Evaluación del cuarto árbol

En la figura 6.4 se presenta el árbol de clasificación con cinco nodos terminales.

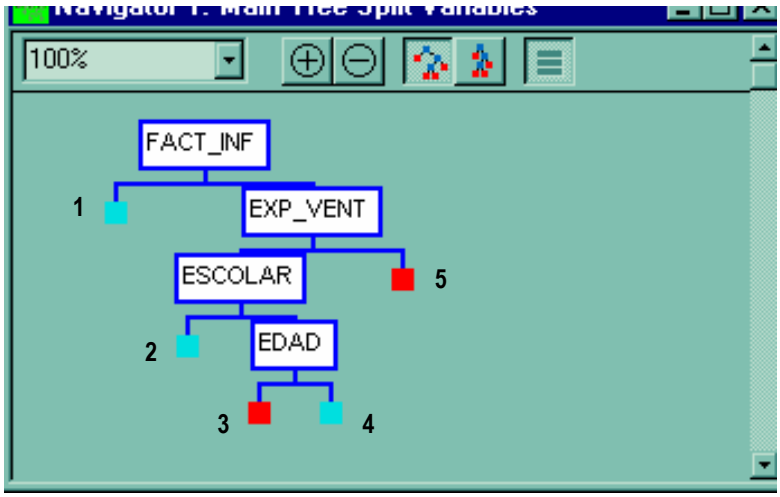


Figura 6.4 Cuarto árbol de clasificación

CART identifica como “el mejor” nodo terminal al número cinco. La tabla 6.15 presenta un resumen del número de casos en los nodos terminales.

Tabla 6.15 Resumen de nodos terminales

Color de nodo	Nodo terminal	Casos por nodo	% de contribución en el nodo	Pureza del nodo
rojo	**5**	**50**	**84.74**	Nodos con mayor pureza o con el mayor número de casos que pertenecen a la clase uno
	3	**26**	**65**	
azul	2	6	16.66	Nodos con pobre concentración de casos de la clase uno
	4	1	16.66	
	1	1	1.69	

Importancia de las variables en la obtención del liderazgo en un año

La tabla 6.16 proporciona la importancia que tiene cada variable predictora en la construcción del árbol. El índice Gini asignó la puntuación a cada variable para hacer las divisiones correspondientes en el árbol.

Tabla 6.16 Importancia de cada variable predictora

Variable	Puntuación	Descripción
FACT_INF	100.00	Factores positivos o negativos que influyeron en la toma de decisión
EDO_CIV	70.77	Estado civil de la líder cuando alcanzó este nivel
PRIORNEG	56.53	Prioridad del negocio para la líder
EXP_VENT	38.25	Evaluación de la experiencia en ventas de la líder
EDAD	32.68	Edad de la líder cuando alcanzó este nivel
ESCOLAR	24.99	Grado máximo de escolaridad de la líder cuando alcanzó este nivel
EMPLEO_2	18.77	Empleo adicional de la líder
TOMAR_DE	10.90	Evaluación de la toma de decisión para ser líder en un año
PERS_INF	7.75	Persona que influyó determinantemente en la toma de decisión
GUSTCIA	5.37	Factor de la compañía K de mayor agrado para la líder
MOTDEMP	2.16	Motivo principal para dejar el empleo adicional
HIJOS	0.00	Evaluación de los hijos que son dependientes económicos de la líder
EXPCIA	0.00	Evaluación de las expectativas de la líder sobre la empresa K

Nuevamente la variable FACT_INF (factores positivos o negativos que influyeron en la toma de decisión), fue la que obtuvo mayor porcentaje en la evaluación del índice Gini y la variable con cero porcentaje fue EXPCIA (evaluación de las expectativas de la líder sobre la empresa K).

Resultados

Reglas de clasificación para las distribuidoras que obtienen el liderazgo de ventas en un año

El principal patrón que siguen estas distribuidoras es:

Nodo 5: Necesidad económica, independencia laboral, confianza en la compañía, los hijos, autorrealización, tener un cambio de vida, reconocimiento profesional, ser ejecutiva independiente y que la mujer tenga experiencia en ventas. Ver figura 6.5.

Reglas de clasificación para las distribuidoras que no obtienen el liderazgo de ventas en un año

Nodo 1: La distribuidora presenta alguna o varias de las siguientes características: timidez, pareja en desacuerdo, interés único por las ventas, trabajo alterno absorbente, desconocimiento de los beneficios, sin experiencia en ventas, falta de decisión, falta de disciplina, falta de capacitación y no tener quien cuide a sus hijos.

Nodo 2: Las características anteriores del nodo uno y además que la mujer no tiene experiencia en ventas y su instrucción escolar es baja o media como la primaria, la secundaria o carrera técnica; hace probable que no obtenga el liderazgo. Ver figura 6.5.

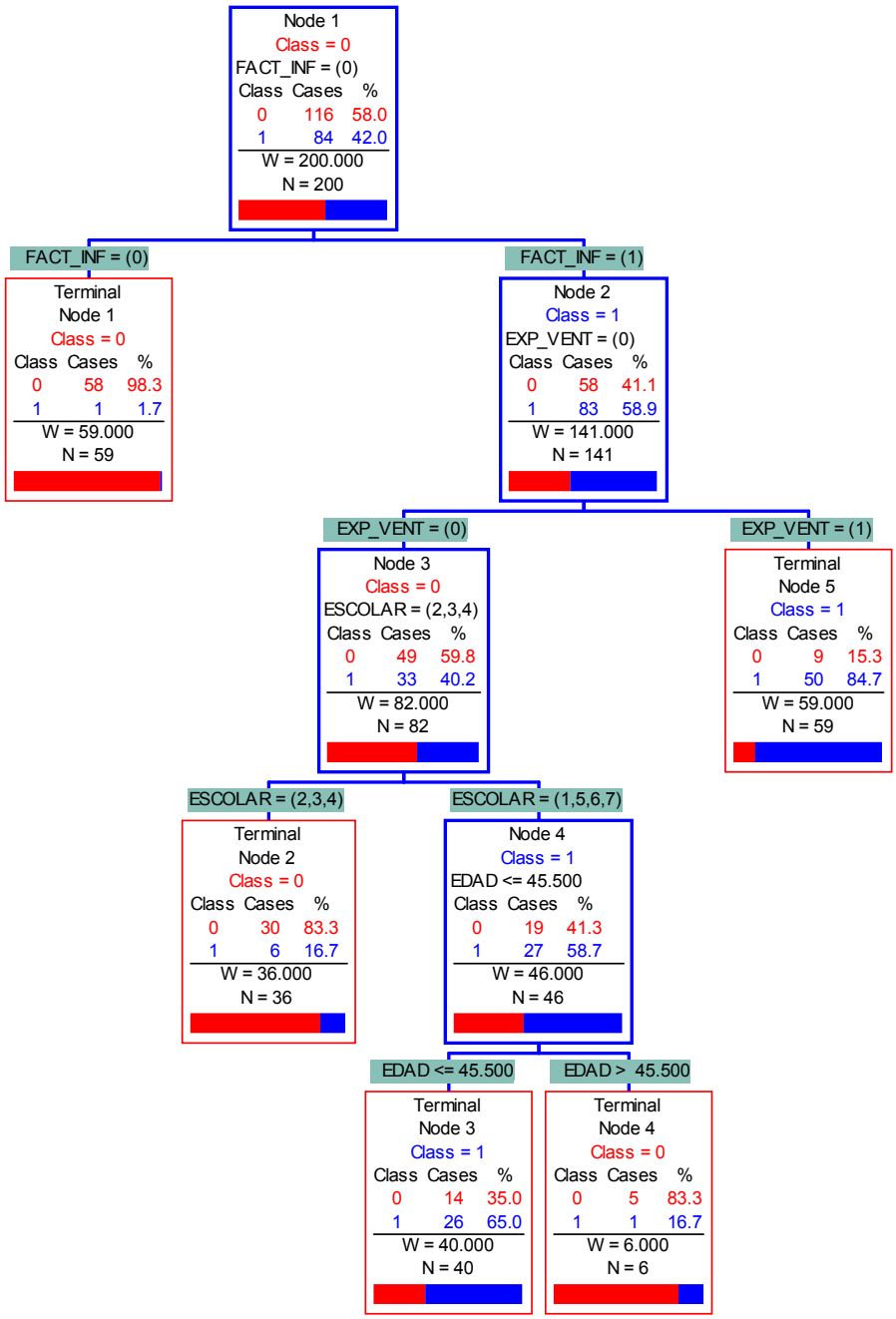


Figura 6.5 *Cuarto árbol con especificaciones*

Validación de resultados

En este modelo, el algoritmo también divide el conjunto de datos de forma aleatoria en dos conjuntos de datos, el de aprendizaje y el de validación.

Evaluando la capacidad del algoritmo CART en la muestra de aprendizaje, se observa en la Tabla 6.17 que la proporción de casos clasificados erróneamente fue 8 de 80 de la clase uno representando el 10% y 23 de 120 de la clase cero representando un 19.16%. Esto da como resultado un porcentaje de clasificación correcta del 84.5% del total de las distribuidoras que se iniciaron en este negocio y lograron el liderazgo de ventas.

Tabla 6.17 Clasificaciones correctas e incorrectas en la muestra de aprendizaje en el cuarto modelo

Muestra de aprendizaje						
Clase	**Casos**	**Casos clasificados correctamente**	**% de clasificación correcta**	**Casos clasificados incorrectamente**	**% de clasificación incorrecta**	**Casos predichos**
1	80	72	90	8	10	99
0	120	97	80.83	23	19.16	101
Total	200	169	84.5	31	15.5	200

Evaluando la muestra de validación, se observa en la Tabla 6.18 que la proporción de casos clasificados erróneamente fue 12 de 80 de la clase uno representando el 15 % y 33 de 120 de la clase cero representando un 27.5%. Esto da como resultado un porcentaje de clasificación correcta del 77.50% del total de las distribuidoras que se iniciaron en este negocio y lograron el liderazgo de ventas.

Conclusión: Comparando las muestras de validación y la de aprendizaje el porcentaje de clasificación correcta de la de validación es más bajo, lo que significa que el modelo aprendido se comporta de forma general y no de forma específica con los datos de prueba, para verificarlo se realizará el procedimiento de poda en este modelo y comparar resultados.

Tabla 6.18 Clasificaciones correctas e incorrectas en la muestra de validación del cuarto modelo

Muestra de validación						
Clase	**Casos**	**Casos clasificados correctamente**	**% de clasificación correcta**	**Casos clasificados incorrectamente**	**% de clasificación incorrecta**	**Casos predichos**
1	80	68	85	12	15	105
0	120	87	72.5	33	27.5	95
Total	200	155	77.50	45	22.5	200

6.6 Evaluación del quinto modelo: estrategia de poda aplicada al cuarto modelo

Al realizar la estrategia de poda en el cuarto modelo, la rama eliminada es EDAD (Edad de la líder) y los nodos terminales eliminados son tres y cuatro; el nodo cinco, que es considerado por el algoritmo como “el mejor” queda igual. Las ramas conservadas son FACT_INF (Factores Positivos o negativos que influyeron en la toma de decisión), EXP_VENT (Experiencia en ventas) y ESCOLAR (Grado máximo de escolaridad). (Ver figuras 6.6 (a) y 6.6 (b)).

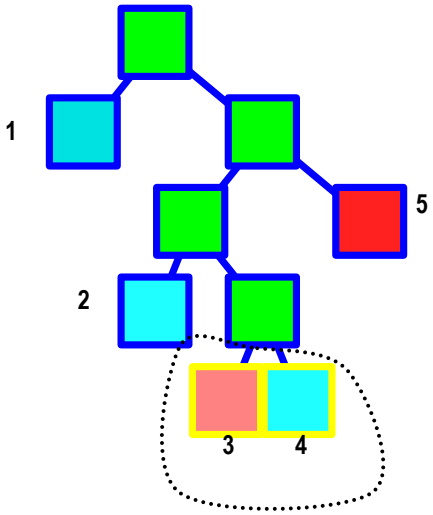


Figura 6.6 (a) *Estrategia de poda en el cuarto modelo*

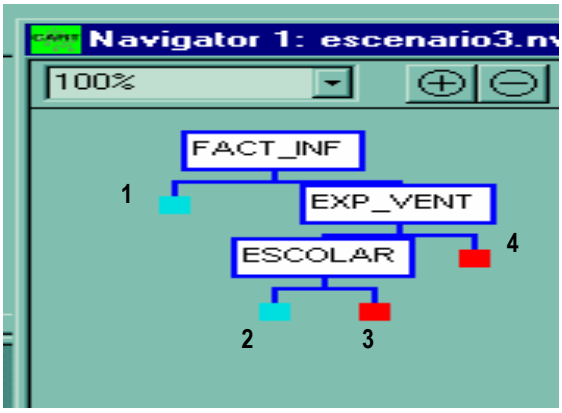


Figura 6.6 (b) *Modelo resultante de la primera poda*

Comparación de nodos terminales con mayor contribución para la clase uno en ambos modelos

En la Tabla 6.19 se observa que, tanto en el modelo original como en el de poda, se conservó el porcentaje de casos de la clase 1 en el nodo considerado como "el mejor" (ahora nodo cuatro), ya que la única variable afectada fue la de EDAD.

Tabla 6.19 Comparación de nodos terminales en ambos modelos

Modelo original		**Modelo con una poda**	
Nodo	**Contribución**	**Nodo**	**Contribución**
5	**84.74%**	**4**	**84.746%**
3	65%	3	58.696%
2	16.66%	2	16.667%
4	16.66%	1	1.695%
1	1.69%		

Comparación de las variables con mayor puntaje para la ramificación

En la Tabla 6.20 se observa que, tanto en el modelo original como en el de poda, el algoritmo tomó la mayoría de las variables, excluyendo únicamente las variables HIJOS y EXPCIA en ambos modelos.

Tabla 6.20 Comparación de las variables predictoras en ambos modelos

	Modelo original		Modelo con una poda	
	Variable	**Puntuación**	**Variable**	**Puntuación**
1	FACT_INF	100.00	FACT_INF	100.00
2	EDO_CIV	70.77	EDO_CIV	63.10
3	PRIORNEG	56.53	PRIORNEG	56.53
4	EXP_VENT	38.25	EXP_VENT	38.25
5	EDAD	32.68	ESCOLAR	24.99
6	ESCOLAR	24.99	EDAD	24.79
7	EMPLEO_2	18.77	EMPLEO_2	18.77
8	TOMAR_DE	10.90	TOMAR_DE	10.94
9	PERS_INF	7.75	PERS_INF	7.44
10	GUSTCIA	5.37	GUSTCIA	5.37
11	MOTDEMP	2.16	MOTDEMP	2.16
12	HIJOS	0.00	HIJOS	0.00
13	EXPCIA	0.00	EXPCIA	0.00

Resultados

Las reglas de clasificación tanto para las distribuidoras que obtienen el liderazgo de ventas en un año como para las que no, son las mismas que en el modelo original (antes de aplicar la estrategia de poda).

Validación de resultados

Muestra de aprendizaje

Las tablas 6.21 y 6.22 muestran un comparativo de las clasificaciones erróneas en la *muestra de aprendizaje* tanto en el modelo original como en el modelo con poda.

Tabla 6.21 Comparación de las clasificaciones correctas e incorrectas del cuarto modelo y del modelo con poda en la muestra de aprendizaje

Muestra de Aprendizaje						
	Modelo original			Modelo con poda		
Clase	**Casos clasificados correctamente**	**Casos clasificados incorrectamente**	**Casos predichos**	**Casos clasificados correctamente**	**Casos clasificados incorrectamente**	**Casos predichos**
Clase 1 N=80	72	8	99	73	7	105
Clase 0 N=120	97	23	101	92	28	95
Total	176	24	200	165	35	200

Tabla 6.22 Comparación de los porcentajes de las clasificaciones correctas e incorrectas del cuarto modelo y del modelo con poda en la muestra de aprendizaje

Muestra de Aprendizaje				
	Modelo original		**Modelo con poda**	
Clase	**% de clasificación correcta**	**% de clasificación incorrecta**	**% de clasificación correcta**	**% de clasificación incorrecta**
Clase 1 N=80	90	10	91.25	8.75
Clase 0 N=120	80.83	19.16	76.66	23.33
Total	84.5	15.5	82.50	17.50

En el modelo con poda, la proporción de casos mal clasificados en la clase uno, fue 7 de 80, representando el 8.75% y en la clase cero fue 28 de 120, representando el 23.33%.

Comparado con el modelo original la clase uno varia su porcentaje de clasificación incorrecta en 1.25% y en la clase cero varía por 4.14%, siendo más alto el porcentaje de error en el modelo con poda en la clase cero. El porcentaje de clasificación correcta en el modelo original es de 84.5% mientras que en el de poda es de 82.50%.

Muestra de validación

Las tablas 6.23 y 6.24 muestran un comparativo de las clasificaciones erróneas en la *muestra de validación* tanto en el modelo original como en el modelo con poda.

Tabla 6.23 Comparación de las clasificaciones correctas e incorrectas del cuarto modelo y del modelo con poda en la muestra de validación

Muestra de Validación					
	Modelo Original			Modelo con Poda	
Clase	**Casos clasificados correctamente**	**Casos clasificados incorrectamente**	**Casos predichos**	**Casos clasificados correctamente**	**Casos clasificados incorrectamente**	**Casos predichos**
Clase 1 N=80	68	12	105	67	13	105
Clase 0 N=120	87	33	95	86	34	95
Total	155	45	200	153	47	200

Tabla 6.24 Comparación de los porcentajes de las clasificaciones correctas e incorrectas del cuarto modelo y del modelo con poda en la muestra de validación

Muestra de Validación				
	Modelo original		Modelo con poda	
Clase	**% de clasificación correcta**	**% de clasificación incorrecta**	**% de clasificación correcta**	**% de clasificación incorrecta**
Clase 1 N=80	85	15	83.75	16.25
Clase 0 N=120	72..50	27.50	71.66	28.33
Total	77.50	22.5	76.5	23.5

En el modelo con poda, la proporción de casos mal clasificados en la clase uno, fue 13 de 80, representando el 16.25% y en la clase cero fue 34 de 120, representando el 28.33%. Comparado con el modelo original la clase uno varia su porcentaje de clasificación incorrecta en 1.25% y en la clase cero varía por 0.83%, siendo más alto el porcentaje de error en el modelo con poda. El total de clasificación correcta en el modelo original es de 77.5% y en el de poda es de 76.5%

Conclusión: Se observa que el comportamiento del modelo tanto en el original como en el de poda es muy semejante en ambas muestras (aprendizaje y validación), esto significa que en este caso la estrategia de poda no afectó el comportamiento del modelo.

6.7 Análisis de resultados

La Tabla 6.25 presenta un comparativo de los resultados entre los cinco modelos. El primer modelo se utilizó para un análisis preliminar de los datos, los porcentajes de clasificación son muy aceptables, pero no se puede tomar como modelo predictivo ya que el ajuste es muy específico al conjunto de datos examinado. Los demás modelos, obtuvieron *porcentajes aceptables de clasificación correcta*, pero de acuerdo a la muestra de validación, el cuarto modelo o árbol generado es el que mejor resultados arroja con un porcentaje de clasificación correcta del 77.50% y un porcentaje de clasificación incorrecta del 22.5%, esto indica que el algoritmo CART obtenido tiene poder predictivo.

Tabla 6.25 Comparativo de resultados entre modelos

Modelo	**Muestra**	**% clasificaciones correctas**	**% clasificaciones erróneas**
Primero	Aprendizaje	92	8
Segundo	Aprendizaje	88	12
	Validación	77	23
Tercero	Aprendizaje	82.50	17.50
	Validación	76	24
Cuarto	Aprendizaje	84.5	15.5
	Validación	**77.50**	**22.5**
Quinto	Aprendizaje	82.50	17.50
	Validación	76.5	23.5

La Tabla 6.26 presenta un extracto del archivo de salida, la columna *PREDECIDO* muestra la predicción del algoritmo para la distribuidora (obtuvo o no el liderazgo de ventas) y la columna *TIPOLIDE,* muestra los valores iniciales de la investigación para realizar comparaciones.

Tabla 6.26 Extracto del archivo de salida

ID_LIDER	FACT_INF	HIJOS	TOMAR_DE	PERS_INF	EXP_VENT	EMPLEO_2	MOTDEMP	PRIORNEG	EDAD	ESCOLAR	EDO_CIV	GUSTCIA	EXPCIA	TIPOLIDE	PREDECIDO
122	1	1	0	4	1	4	1	1	28	5	1	2	0	1	1
124	1	1	0	4	1	5	1	2	34	4	1	3	1	1	1
125	1	1	0	2	1	3	1	1	36	4	2	6	0	1	0
126	1	0	0	4	1	5	1	1	34	5	2	3	1	1	0
129	1	1	0	4	1	1	2	1	30	6	1	1	1	1	0
131	1	0	1	3	1	5	4	1	28	5	1	3	1	1	1
136	1	1	1	3	1	7	5	1	55	6	2	1	0	1	0
140	1	0	0	3	1	3	1	1	33	4	2	6	3	1	0
141	1	1	0	1	1	1	1	1	27	6	1	6	1	1	0
142	1	1	0	4	1	4	1	1	30	5	2	5	1	1	0
144	1	0	0	3	1	3	1	1	45	5	3	6	1	1	0
147	1	1	0	1	1	4	2	1	32	5	1	4	1	1	0
151	0	1	0	1	0	3	2	2	25	6	1	6	1	1	0
6	1	1	0	2	1	4	1	1	39	4	2	4	0	0	1
13	1	1	0	1	1	3	1	1	40	4	2	6	1	0	1
17	1	1	1	2	0	3	2	2	45	4	2	5	0	0	1
18	1	0	1	1	0	6	1	2	45	2	2	6	1	0	0
19	1	1	1	2	0	3	2	1	28	5	2	5	0	0	0
20	1	1	1	3	0	7	2	1	46	6	2	1	1	0	0
25	1	0	1	2	0	5	5	2	45	4	2	4	0	0	0
27	1	0	1	1	0	6	2	2	36	3	5	4	0	0	0
29	1	1	1	3	0	5	4	2	43	5	2	6	0	0	0

Patrones de comportamiento

El patrón de comportamiento según el modelo para las distribuidoras que obtienen el liderazgo de ventas en un año es:

Son mujeres con experiencia en el área de ventas, tienen necesidad económica, les gusta tener independencia laboral, tienen confianza en la compañía, sus hijos son su principal motivación, desean auto-realizarse profesionalmente, desean un cambio en su vida y les gusta ser reconocidas por su trabajo a su trabajo.

Por el contrario las mujeres que no llegan al liderazgo son tímidas, su pareja está en desacuerdo con su actividad que desempeñan en la compañía, su único interés en la compañía son las ventas y no les atrae la idea de desarrollar su grupo de compra, algunas de ellas tienen trabajos alternos de tiempo completo, algunas desconocen los beneficios que tiene el liderazgo de ventas, otras no tienen experiencia en ventas, son mujeres que no tienen poder de decisión y algunas su instrucción escolar es baja o media.

Medidas correctivas y control de mejoras

Los resultados obtenidos parten de la premisa de que las distribuidoras ya fueron evaluadas previamente con un cuestionario de reconocimiento. Además el poder predictivo de los resultados del algoritmo CART supone que el comportamiento de las futuras distribuidoras tendrá las mismas características de las distribuidoras evaluadas, características con las cuales se obtuvieron los resultados.

Si las condiciones cambian, es decir, se identifican cambios de comportamiento en las distribuidoras, entonces los resultados del algoritmo CART no funcionarán bien, y en todo caso se deberá aplicar un nuevo algoritmo.

Acciones sugeridas

Para obtener resultados aceptables del *árbol de clasificación* generado se sugiere el siguiente procedimiento.

En cada línea de auspicio, se debe:

1. Aplicar a cada nueva distribuidora un cuestionario de reconocimiento"[34], para determinar sus características.
2. Recopilar mensualmente en cada línea de auspicio, los cuestionarios aplicados.
3. Vaciar la información codificada a un archivo.
4. Aplicar el modelo a la información del archivo anterior.
5. Regresar la información correspondiente a cada líder de la línea de auspicio, con los resultados obtenidos en el modelo, con la finalidad de que ella identifique las áreas de oportunidad de cada nueva distribuidora, así como a las posibles candidatas a ser líderes de venta en un año. Este punto es muy importante, ya que aunque se identifique a elementos candidatos a ser líderes, si no se tiene un seguimiento adecuado y constante, esta metodología no servirá.

Una vez, realizada la identificación, se podrán seguir estrategias de mercadotecnia en conjunto con la empresa para dar empuje a estas distribuidoras. El seguimiento se dará de forma mensual, y posteriormente podrán obtenerse otros análisis, como por ejemplo, identificar períodos de mayor reclutamiento, identificación del porcentaje de líderes de venta por zonas, entre otros.

Al implementarse el modelo, cada vez se realizarán los ajustes necesarios, debido a que se encontrarán nuevas variables y otras que serán innecesarias.

[34] Ver Apéndice B.

Conclusiones

La dinámica de crecimiento del sistema de venta directa es constante y muy rápida, favoreciendo la creación de importantes fuentes de trabajo que ayudan a las economías de los países. Alrededor de la venta directa, se gestan interesantes temas de investigación en diferentes áreas, como en este caso en particular, que el emplear la técnica de los árboles de clasificación sirvió para encontrar el patrón de comportamiento que siguen las distribuidoras de la compañía K para obtener el nivel de liderazgo de ventas en un año.

Así mismo, utilizando el algoritmo CART se ha conseguido detectar las variables que más influyen en las distribuidoras para obtener el liderazgo de ventas, éstas variables son: los factores tanto positivos como negativos que influyen en la mujer para la toma de la decisión de ser líder, la experiencia en el área de ventas, el grado máximo de escolaridad y la edad. El modelo es un árbol con cinco nodos terminales, fácil de interpretar, comprensible e inteligible, que clasificó correctamente 155 casos de un total de 200 consiguiendo un porcentaje de clasificación correcta del 77.50%, lo cual para la compañía K es un resultado muy aceptable.

Este modelo es predictivo ya que se utilizará en nuevos conjuntos de datos de distribuidoras, previamente evaluadas con un cuestionario de reconocimiento, para detectar aquellas mujeres que tienen mayor posibilidad de ser líderes en un año.

La capacidad de pronóstico del algoritmo aumenta con el número de características disponibles (variables independientes o explicativas) que se utilicen en el cuestionario de reconocimiento.

Lo anterior confirma la hipótesis sustentada “el árbol de clasificación generado por el algoritmo CART identifica rápidamente el patrón de comportamiento existente en las distribuidoras que obtienen el liderazgo de ventas en un año, teniendo el poder predictivo para utilizar estas reglas y clasificar eficazmente a grupos futuros de nuevas distribuidoras, con un porcentaje mínimo de error”.

Usando esta metodología de trabajo se ha automatizado el proceso de clasificación de distribuidoras, proceso que actualmente se realiza por intuición de las líderes de grupo, logrando con mayor precisión si una mujeres es candidata a líder de ventas.

Además bajo este análisis de información, la compañía K puede tener proyecciones más seguras y en tiempo real de quienes son sus mejores prospectos a líderes.

Como trabajos de seguimiento y adicionales a ésta investigación, se pueden hacer comparaciones de los resultados obtenidos con el algoritmo CART con resultados en otros algoritmos de árboles de decisión como por ejemplo el CHAID (detector automático de interacciones de chi-cuadrado).

El modelo generado debe ser implementado y evaluado a mediano y largo plazo, con la finalidad de realizar los ajustes necesarios en los parámetros y en la construcción del mismo para seguir alcanzando resultados óptimos y certeros.

Como se observó la minería de datos, aporta diariamente nuevos cursos de acción para afrontar diversas problemáticas en distintas áreas de las organizaciones, además con el fortalecimiento de la tecnología, los profesionales en el ramo pueden implementar innovadoras y factibles soluciones al surgimiento de nuevos problemas.

APÉNDICE A

Se presenta el cuestionario de reconocimiento aplicado a las actuales líderes de ventas de la compañía K, con la finalidad de identificar las características que las llevaron a obtener este nivel en la compañía en un tiempo determinado.

Folio:_____

CUESTIONARIO DE RECONOCIMIENTO PARA LIDERES DE VENTAS

Para la empresa K es prioritario que sus distribuidoras independientes alcancen el nivel de líder de ventas en un tiempo óptimo, es por ello que se realiza un estudio para detectar los factores principales que influyeron en las actuales líderes de ventas para lograr este nivel en la compañía.

Instrucciones de llenado.

a) Lee cuidadosamente cada pregunta.

b) Cuando la pregunta sea abierta, contesta en forma breve.

c) En las preguntas que tengan opciones marca con una X la respuesta que para ti sea la indicada.

d) No dejes casillas vacías.

Nota: esta información es de carácter confidencial. Gracias por tu colaboración.

1.- Indica el tiempo que llevas en la compañía: _____años.

2.- Indica el tiempo que tienes de ser líder de ventas: _____.

3.- ¿Tenías hijos en edad escolar, cuando llegaste al nivel de líder de ventas?

a) _____SÍ

b) _____NO

4.- Indica, ¿qué factor influyó de manera significativa para que lo hicieras en este tiempo?_____

1) Necesidad económica
2) Independencia laboral
3) Confianza en la compañía
4) Los hijos
5) Autorrealización
6) Un cambio de vida
7) Reconocimiento profesional
8) Ser una ejecutiva independiente
9) Timidez
10) Pareja en desacuerdo
11) Interés único por las ventas
12) No había quien cuidara a los hijos
13) Trabajo adicional absorbente
14) Desconocimiento de los beneficios
15) Sin experiencia
16) Falta de decisión propia
17) Por falta de disciplina
18) Por falta de capacitación

5.-¿Te costó trabajó tomar la decisión de ser líder de ventas?

a) _____SÍ

b) _____NO

6.-¿Qué persona influyó de manera determinante para que tomaras la decisión?

a)_____ Líder de grupo

b) _____Hijos

c) _____ Familia

d) _____Yo misma

e)_____Otra

7.- ¿Tenías experiencia en actividades de ventas o negocios?

a) _____SÍ

b) _____NO

8.- Antes de ser líder de ventas, ¿cuál era tu ocupación?

a)_____Profesional o Técnica

b) _____Gerentes o Administradoras

c) _____Ama de casa

d) _____Vendedoras

e)_____Trabajadora de oficina

f) _____Artesanas / Obreras / Limpieza

g) _____Sector Educativo

h) _____Otro

9.-¿Qué te motivó a dejar tu empleo anterior?

a)_____Mayor remuneración económica

b) _____Mayor reconocimiento

c) _____Mayor capacitación

d) _____Un horario flexible

e)_____Independencia laboral

10.- Indica, ¿cuál era la prioridad de tu negocio?

a) _____Alta

b) _____Media

c) _____Baja

11.- ¿Cuál era tu edad cuando llegaste a ser líder de ventas? _____

12.- ¿Cuál era tu grado de escolaridad?

a) _____Ninguna

b) _____Primaria

c) _____Secundaria

d) _____Técnico o Comercio

e) _____Bachillerato

f) _____Licenciatura

g) _____Postgrado

13.- ¿Cuál era tu estado civil?

a) _____Soltera

b) _____Casada

c) _____Viuda

d) _____Divorciada o Separada

e) _____Unión Libre

14 ¿Qué es lo que más te gusta de la compañía?

a) _____Filosofía y misión de la compañía

b) _____Ganancias económicas

c) _____Gente

d) _____Productos

e) _____Reconocimientos

f) _____Negocio independiente

15.-En el tiempo que tienes como líder de ventas, ¿la compañía ha cumplido tus expectativas personales?

a) _____Siempre

b) _____Casi siempre

16. Actualmente, ¿cuántas líderes de venta forman tu grupo de compra?_____

17.- ¿Qué factores consideras que son los principales para que una mujer no desarrolle su carrera ejecutiva en un tiempo óptimo?

a) _____Miedo

b) _____Hijos

c) _____Pareja

d) _____Otro empleo

e) _____Desconocimiento

f) _____Falta de capacitación

g)_____Impedimento físico

18.- Indica ¿cuántas líderes de venta te gustaría desarrollar en tu grupo de compra por año?_____

APÉNDICE B

Se presenta el cuestionario de reconocimiento que debe aplicarse a cada nueva distribuidora que ingrese a la compañía a través de alguna línea de auspicio, con la finalidad de formar el almacén de datos que alimentará al modelo.

Folio:_____

CUESTIONARIO PARA NUEVAS DISTRIBUIDORAS

Instrucciones de llenado.

a) Lee cuidadosamente cada pregunta.

b) En las preguntas que tengan opciones marca con una X la respuesta que para ti sea la indicada.

c) No dejes casillas vacías.

Nota: esta información es de carácter confidencial. Gracias por tu colaboración.

1.- Indica tu edad: _____ años

2.- Tu Estado Civil es:

a) _____Soltera

b) _____Casada

c) _____Viuda

d) _____Divorciada o Separada

e) _____Unión libre

3.- Tu grado de escolaridad es:

a) _____Ninguna

b) _____Primaria

c) _____Secundaria

d) _____Técnico o Comercio

e) _____Bachillerato

f) _____Licenciatura

g) ____Postgrado

4.- ¿Tienes hijos que dependan económicamente de ti?

a) _____SÍ

b) _____NO

5.- Indica tu actividad principal con la que compartes tu negocio.

a)_____ Profesional o Técnica

b) _____ Gerentes o Administradores

c) _____ Ama de casa

d) _____ Vendedores

e)_____ Trabajadora de oficina

f) _____ Artesanas / Obreras / Limpieza

g) _____ Sector Educativo

h) _____ Otro

6.- Si pudieras modificar alguna situación de tu actividad mencionada en la pregunta anterior, ¿cuál sería la más importante para ti?. Marca sólo una opción.

a) Mejor remuneración económica ( )

b) Mayor reconocimiento o motivación ( )

c) Mayor capacitación ( )

d) Horario flexible ( )

e) Independencia laboral ( )

7.- ¿Tienes experiencia en el área de ventas?

a) _____SÍ

b) _____NO

8.- Selecciona ,¿Cuál fue la principal causa para entrar a formar parte de la compañía?

a) _____Ganar dinero

b) _____Satisfacción personal o autorrealización

c) _____Reconocimiento

d) _____Pertenencia a un grupo

e) _____Independencia laboral

9.- ¿Qué es lo que más te gusta de la compañía?. Marca sólo una opción.

a) _____La filosofía y misión de la compañía

b) _____Las ganancias económicas

c) _____Su gente

d) _____Los productos

e) _____Los reconocimientos como: premios, viajes, autos, joyas, etc.

f) _____Negocio independiente

10.- ¿Qué prioridad asignas a tu negocio como distribuidora independiente?

a) _____Alta

b) _____Media

c) _____Baja

11.- ¿Conoces todos los beneficios que ofrece el nivel de líder de ventas?

a) _____SÍ

b) _____NO

12.- ¿Qué personaje consideras que influye sobre ti de manera determinante para que tomaras la decisión de ser líder de ventas?

a) Líder de grupo ( )

b) Hijos ( )

c) Familia (pareja padres, hermanos) ( )

d) Tu misma ( )

e) Otra ( )

13.-¿Te gustaría ser líder de ventas en el lapso de un año?

a) _____SÍ (Pase a la pregunta 14)

b) _____NO (Pase a la pregunta 15)

14.-¿Qué factor sería el que te motivaría significativamente a tomar la decisión de ser líder de ventas?_____

a) Necesidad económica

b) Independencia Laboral

c) Confianza en la compañía

d) Los hijos

e) Autorrealización

f) Un cambio significativo en la vida

g) Reconocimiento profesional

h) Ser una ejecutiva independiente

i) Otra

15.-¿Qué factor consideras que sería el que te impediría tomar la decisión de ser líder de ventas?_____

a) Pareja en desacuerdo

b) Únicamente me interesan las ventas

c) No tengo con quien dejar a los hijos

d) Mi trabajo adicional me absorbe

e) No conozco los beneficios

f) No tengo experiencia

g) Por falta de decisión propia

h) Por falta de capacitación

i) Otra

BIBLIOGRAFÍA

1. BREIMAN, LEO & FRIEDMAN, JEROME H. (1984). *Classification and Regression Trees.* Chapman & Hall.
2. Cámara Nacional de la Industria de la Perfumería y Cosmética. (2003). *Memoria Estadística CANIPEC.* México, D.F.
3. CLEMENTS, LEONARD W. (189). *Mitos y Verdades del Negocio de Multinivel.* México, D.F., 1998. Edit. Panorama.
4. ESCALANTE, BEATRIZ. (2006). *Curso de Redacción para escritores y periodistas.* 9ª. Edición. México.D.F. Porrúa.
5. ETZIONI, O. (1996). *The World Wide Web:quagmire or gold mine?.* Communications of the ACM.
6. FAYYAD, U.M., PIATETSKY-SHAPIRO, G. & SMYTH, P. (1996). *The KDD Process For Extracting Useful Knowledge from Volumes of Data.* Communications of the ACM.
7. HAN, JIAWEI & KAMBER, MICHELINE. (2001). *Data Mining:Concepts and Techniques.* Morgan Kaufmann Publishers, Inc.
8. HAND, DAVID. J., MANNILA, HEIKKI. & SMYTH PADHRAIC. (2001). *Principles of Data Mining.* (Adaptive Computation and Machine Learning Series). MIT, Press.
9. HERNÁNDEZ ORALLO, J., RAMÍREZ QUINTANA, M.J & FERRI RAMÍREZ, C. (2004*). Introducción a la Minería de Datos.* Madrid, España. Pearson Educación.

10. HERNADEZ SAMPIERI, R., FERNÁNDEZ COLLADO, C. & BAPTISTA LUCIO, P. (2003*). Metodología de la Investigación.* México. McGraw-Hill Interamericana.

11. KING, CHARLES W. & ROBINSON JAMES W. (2004). *Los Nuevos Profesionales.* Buenos Aires, Argentina. Time & Money Network Editions.

12. LAGGOS, KEITH B. (1998). *Direct Sales: an Overview.* United States. Donelley.

13. MALHOTRA, NARESH K. (1997). *Investigación de Mercados, Un enfoque práctico.* México, D.F. Prentice-Hall Hispanoamericana, S.A.

14. *Manual de estilo de publicaciones de la American Psychological Association.* (2002). 2ª. Edición. México. D.F. Manual Moderno.

15. MARY KAY ASH. (2003). *Ocurren los milagros.* México. Mry Kay Inc.

16. MENDENHALL, W., SCHEAFFER R. & WACKERLY D. (1986). *Estadística Matemática con Aplicaciones.* México, D.F. Grupo Editorial Iberoamérica.

17. PEREZ, CESAR.(2001). *Técnicas Estadísticas con SPSS.* Madrid, España. Pearson Educación.

18. POE, RICHARD. (1998). *Cómo Formar su Línea de Auspicio en Multinivel*. México, D.F. Panorama.

19. PYLE, DORIAN (1999). *Data Preparation for Data Mining with CD ROM.* San Fco., Cal. Morgan Kaufmann Publishers, Inc.

20. SALFORD SYSTEMS.(2002). *CART for Windows. User´s Guide.* San Diego, USA.

21. The Direct Selling Education Foundation (U.S.A). (1997). Asociación Mexicana de Ventas Directas, A.C. *Memorias del Primer Seminario Académico.* México, D.F.

22. The Direct Selling Education Foundation (U.S.A). (2000). Cámara Costarricense de Empresas de Venta Directa. *Memorias del Primer Congreso Académico Regional de Ventas Directas.* San José, Costa Rica.

23. QUINLAN. *C4.5. Programs for Machine Learning.* (1993). San Fco., Cal. Morgan Kaufmann Publishers, Inc.

REFERENCIAS ELECTRÓNICAS

24. American Association for Artificial Intelligence (AAAI). http://www.aaai.org/

25. Árboles.http://www-etsi2.ugr.es/depar/ccia/rf/www/tema3_00-01_www/node25.html

26. Árboles.http://www-etsi2.ugr.es/depar/ccia/rf/www/tema3_00-1_www/node26.html

27. Árboles.http://www-etsi2.ugr.es/depar/ccia/rf/www/tema3_00-1_www/node27.html

28. Árboles.http://www-etsi2.ugr.es/depar/ccia/rf/www/tema3_00-1_www/node28.html

29. Árboles.http://www-etsi2.ugr.es/depar/ccia/rf/www/tema3_00-1_www/node29.html

30. Árboles.http://www-etsi2.ugr.es/depar/ccia/rf/www/tema3_00-1_www/node30.html

31. Asociación Mexicana de Venta directa. http://www.amvd.org.mx/amvd/

32. Cámara Nacional de la Industria de la Perfumería, Cosmética y Artículos de Tocador e Higiene. http://www.canipec.org.mx/

33. Diccionario de la Lengua Española. Vigésima segunda edición. http://buscon.rae.es/drael/

34. Enciclopedia Wikipedia. http://es.wikipedia.org/wiki/Coeficiente_Gini

35. Entrepreneur.http://www.soyentrepreneur.com/paina.hts?N=9757&Ad=S

36. Inducción de Árboles de Decisión (TDIDT:Top Down Induction of Decision Trees) http://ccc.inaoep.mx/~emorales/cursoso/Aprendizaje/node7.html

37. SALFORD-SYSTEMS. http://www.salford-systems.com/cart.php