VNIVERSIDAD NACIONAL
AVTNOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

"MINERÍA DE DATOS MULTIPERSPECTIVA"

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN CIENCIAS (COMPUTACIÓN)

P R E S E N T A:

MARÍA DEL ROSARIO CRUZ MARTÍNEZ

DIRECTOR DE TESIS: DR. CHRISTOPHER RHODES STEPHENS STEVENS

México, D.F. **2007**

Agradecimientos

Gracias a mi familia por todo su amor y apoyo incondicional.

Gracias a los amigos con quienes he compartido momentos muy especiales.

Gracias a la UNAM, al IIMAS, y a sus profesores, por contribuir a mi desarrollo académico, profesional y personal.

Gracias al Dr. Christopher Stephens por su guía en la realización de esta tesis y por compartir conmigo su gran experiencia académica y profesional.

Gracias al personal administrativo del IIMAS por la eficiencia y amabilidad que tienen al realizar su trabajo.

También quiero agradecer de manera especial a Adaptive Technologies por todo el apoyo brindado para la realización de esta tesis.

Y a todos aquellos con quienes me he cruzado en el camino y han contribuido a mi crecimiento, muchas gracias.

Tabla de contenido

Introducción

Desde la aparición de las primeras computadoras personales, a principios de los ochentas, ha habido un desarrollo tecnológico impresionante. Actualmente se cuenta con grandes capacidades de adquisición, almacenamiento, procesamiento y difusión de datos. Estos avances tecnológicos han contribuido al surgimiento de nuevas disciplinas. Una de ellas es la minería de datos.

La minería de datos es una disciplina científica relativamente nueva cuyo objetivo general es extraer conocimiento de los datos. Está basada en otras áreas científicas como son la estadística, bases de datos, inteligencia artificial, reconocimiento de patrones y aprendizaje por máquina.

Las aplicaciones, y áreas de aplicación, de la minería de datos son muy variadas. Algunos ejemplos de aplicaciones de minería de datos son:

- La descripción del perfil sociodemográfico de los compradores de cierta póliza de seguro, con la finalidad de atraer a nuevos clientes.

- La predicción de quienes son los candidatos a estudiar en una universidad que tienen mayores probabilidades de terminar sus estudios.

- La predicción de los pacientes con mayores probabilidades de ser muy costosos en el siguiente año dada información del año actual.

- La descripción del perfil de las empresas que tienen mayores probabilidades de adquirir cierto producto.

- Establecer un método efectivo para la búsqueda de imágenes o de texto.

La minería de datos es importante porque permite resolver problemas complejos mediante la exploración y explotación de datos. A pesar de que su uso todavía no está muy extendido, sobre todo en México, poco a poco se va comprendiendo su importancia.

Los problemas reales de minería de datos generalmente son complejos porque no consisten de un objetivo único e involucran toma de decisiones. Cuando una persona tiene que resolver un problema complejo, lo que hace es pedir la opinión de diversos especialistas para poder llegar a una mejor decisión. Lo mismo aplica para la solución de problemas de minería de datos. Lo mejor es seguir un enfoque multiperspectiva que integre distintas opiniones. Este enfoque multiperspectiva se puede tener en distintos niveles, desde el nivel de datos, hasta el nivel de los métodos de minería de datos.

El objetivo de la tesis es mostrar que una metodología multiperspectiva de minería de datos ofrece ventajas significativas contra métodos existentes tradicionales.

Para cumplir con el objetivo planteado se presenta un marco teórico de la minería de datos y se explica en qué consiste el enfoque multiperspectiva. Este enfoque se aplica en la solución de dos problemas de minería de datos: un problema académico y un problema del mundo de los negocios.

No se pretende dar un compendio detallado de los métodos disponibles en la minería de datos, pero sí proporcionar un panorama general de algunos de los métodos más empleados para resolver problemas de minería de datos y dar algunas referencias para profundizar en dichos métodos.

La tesis se divide en cuatro capítulos:

En el primer capítulo se da una introducción a la minería de datos. Esta introducción empieza con un breve recorrido en el surgimiento de esta disciplina y la definición de la minería de datos. Después se describen las tareas que se pueden realizar con la minería de datos y se proporciona un panorama general de algunos de los métodos empleados para realizar dichas tareas. Finalmente, se mencionan algunos aspectos a considerar en la resolución de problemas del mundo real.

En el segundo capítulo se introduce el concepto de "paisaje de predictibilidad". Este concepto es útil para explicar y comprender aspectos importantes en la resolución de problemas de minería datos. También se aborda el problema de la dimensionalidad y se proporcionan formas para afrontar la "maldición de la dimensionalidad". Al final se plantea el enfoque multiperspectiva para la resolución de problemas de minería de datos.

El tercer capítulo es un caso de estudio de un problema de minería de datos planteado en el concurso *Challenge* 2000, el cual consistió en predecir los compradores de una póliza de seguro y en explicar por qué. Si bien este problema es más bien académico, permite ilustrar cómo abordar problemas de minería de datos mediante el enfoque multiperspectiva.

El cuarto capitulo es un caso de estudio de un problema real de minería de datos. Este problema consistió en predecir a los pacientes más costosos del siguiente año con base en información de su condición médica y de sus costos en el año actual. Al igual que en el caso de estudio del capítulo tres, se aborda el problema mediante un enfoque multiperspectiva.

Capítulo 1: ¿Qué es la minería de datos?

1.1 Antecedentes

El acelerado desarrollo tecnológico ha permitido que contemos con dispositivos de adquisición y de procesamiento cada vez más rápidos y con mayores capacidades de almacenamiento a precios bastante accesibles. Esto ha originado que el almacenamiento de datos se incremente día con día. Aún cuando los datos no se hayan recolectado con el fin de analizarlos, las personas son cada vez más conscientes del provecho que se puede obtener al analizar los datos. Todos estos factores han contribuido al surgimiento y desarrollo de una disciplina científica, conocida como minería de datos, la cual se basa en otras áreas científicas, principalmente estadística, bases de datos, inteligencia artificial, aprendizaje por máquina y reconocimiento de patrones. En términos generales, el objetivo de la minería de datos es extraer información de los datos.

El primer taller especializado en minería de datos, KDD-1989, se realizó durante la Onceava Conferencia Internacional Conjunta en Inteligencia Artificial en Detroit, Michigan, en agosto de 1989 (las siglas KDD significan *Knowledge Discovery in Databases*). El taller tuvo una gran respuesta a nivel mundial (se recibieron propuestas de doce países en cuatro continentes) y una gran asistencia, formada por investigadores, así como representantes del gobierno y de la industria [25]. El reporte del taller se puede consultar en la revista de inteligencia artificial *AI Magazine* (volumen 11, número 5).

El interés generado por el taller dio origen a la publicación, en 1991, del primer libro sobre minería de datos [29]: *Knowledge Discovery in Databases* [25]. El libro es una recopilación de los mejores artículos presentados en el taller KDD-1989. Los artículos fueron reescritos para proporcionar un tratamiento uniforme de elementos clave como el uso del conocimiento del dominio en el descubrimiento, la complejidad algorítmica y el manejo de incertidumbre. En el libro se presenta un panorama general del descubrimiento de conocimiento en bases de datos y se estructuran los artículos en siete partes:

1. *Discovery of Quantitative Laws*
2. *Discovery of Qualitative Laws*
3. *Using Knowledge in Discovery*
4. *Data Summarization*
5. *Domain Specific Discovery Methods*
6. *Integrated and Multiparadigm Systems*
7. *Methodology and Application Issues*

Posterior a este taller, se realizaron tres talleres más (1991, 1993 y 1994). Debido al gran interés y estímulo generado por los talleres, éstos evolucionaron en conferencias internacionales anuales, las cuales se han llevado a cabo ininterrumpidamente desde 1994.

A partir de estas conferencias, en 1996 se editó el libro "*Advances in Knowledge Discovery and Data Mining*", el cual se convirtió en una de las principales bases para la investigación en minería de datos.

En noviembre del mismo año, la revista *Communications of the ACM* (vol. 39 No. 11) presentó una sección especial dedicada a la minería de datos. Los artículos presentados en dicha revista fueron:

- *Data Mining and Knowledge Discovery in Databases*
- *The KDD Process for extracting Useful Knowledge from Volumes of Data*
- *Statistical Inference and Data Mining*
- *Mining Business Databases*
- *The Data Warehouse and Data Mining*
- *Mining Scientific Data*
- *A Database Perspective on Knowledge Discovery*
- *The World-Wide Web: Quagmire or Gold Mine?*

En 1997 se creó la primera revista especializada en minería de datos: *Data Mining and Knowledge Discovery Journal*. Esta revista se enfoca en la teoría, técnicas y prácticas para extraer información de grandes bases de datos [26]. Algunos artículos de esta revista se pueden consultar en línea en el sitio de Internet de la editorial Springer.

Al año siguiente se formó el grupo SIGKDD (*Special Interest Group on Knowledge Discovery and Data Mining*), el cual está enfocado en proporcionar el foro más importante para el progreso y el empleo de la ciencia del descubrimiento de información y minería de datos [1]. SIGKDD es uno de los 34 grupos de interés de la ACM (*Association for Computing Machinery*), que junto con la Sociedad de Cómputo de la IEEE (*Institute of Electrical and Electronics Engineers*), son las sociedades más importantes de computación a nivel mundial.

Debido a que el área de minería de datos surgió y se ha desarrollado como un campo interdisciplinario, SIGKDD constantemente interactúa con otras asociaciones de ACM, y externas, como SIGMOD (*SIG on Management of Data*), SIGART (*SIG on Artificial Intelligence*), SIGMIS (*SIG on Management Information Systems*), SIGIR (*SIG on Information Retrieval*), AAAI (*American Association for Artificial Intelligence*), IEEE (*Institute of Electrical and Electronics Engineers*) y ASA (*American Statistical Association*).

El interés en la minería de datos sigue creciendo y cada vez son más las personas interesadas en investigar sobre minería de datos o en aplicarla. Actualmente hay varios congresos relacionados con la minería de datos, algunos de ellos son:

International Conference on Algorithmic Learning Theory,
International Conference on Discovery Science,
Annual Data Mining Technology Conference,
IEEE Symposium on Visual Analytics Science and Technology,
KXEN Extreme Data Mining User Group,
International Conference on Tools with Artificial Intelligence,
Iberoamerican Congress on Pattern Recognition,
Australasian Data Mining Conference,
International Conference on Machine Learning and Applications,
IEEE International Conference on Data Mining,
International Joint Conference on Artificial Intelligence,
Workshop on Analytics for Noisy Unstructured Text Data,
Extraction et Gestion des Connaissances,

IEEE Symposium on Computational Intelligence and Data Mining,
International Conference on Data Engineering,
SIAM International Conference on Data Mining,
ACM SIGMOD-SIGACT-SIGART Symposium on Principles Of Database Systems,
Wessex Int. Conference on Data, Text and Web Mining and their Business Applications,
Industrial Conference on Data Mining,
International Conference on Machine Learning and Data Mining,
Joint Statistical Meetings,
ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,
International Conference on Database and Expert Systems Applications,
European Conference on Machine Learning & European Conference on Principles and Practice of Knowledge Discovery in Databases

1.2 Definición de minería de datos y búsqueda de conocimiento en bases de datos

Existen diversas definiciones de minería de datos, aunque en general coinciden en lo fundamental: la minería de datos es un proceso que extrae información de los datos. Las variaciones en las distintas definiciones se dan principalmente en las características atribuidas a la información que se extrae y al proceso de extracción.

La definición usada en el presente trabajo es la siguiente:

La minería de datos es "la exploración y análisis de los datos para descubrir patrones, correlaciones y otras regularidades" [27].

Un concepto que en ocasiones es considerado como sinónimo de minería de datos es el KDD (*Knowledge Discovery in Databases*).

Algunas personas consideran la minería de datos y el KDD como conceptos sinónimos, mientras que otras consideran que la minería de datos es uno de los pasos que componen al KDD. Fayyad, Piatesky-Shapiro y Smyth son de los últimos. Ellos definen el proceso de KDD como:

"el proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y entendibles en los datos" [10].

Para dejar más clara esta definición, se mostrará un ejemplo sencillo creado a partir de datos reales de un proyecto de minería de datos. Se tienen los valores de dos variables para un conjunto de 33 pacientes diabéticos: edad y gastos totales del paciente en 1997. Se desea usar estas variables para predecir quiénes de estos pacientes tendrán gastos totales mayores a 125,000 dólares en 1998. La siguiente figura muestra un diagrama de dispersión de las dos variables. Los puntos en blanco corresponden a los pacientes que tuvieron gastos sobre los 125,000 dólares en 1998 y los puntos en negro corresponden a los pacientes que tuvieron gastos menores a dicho umbral.

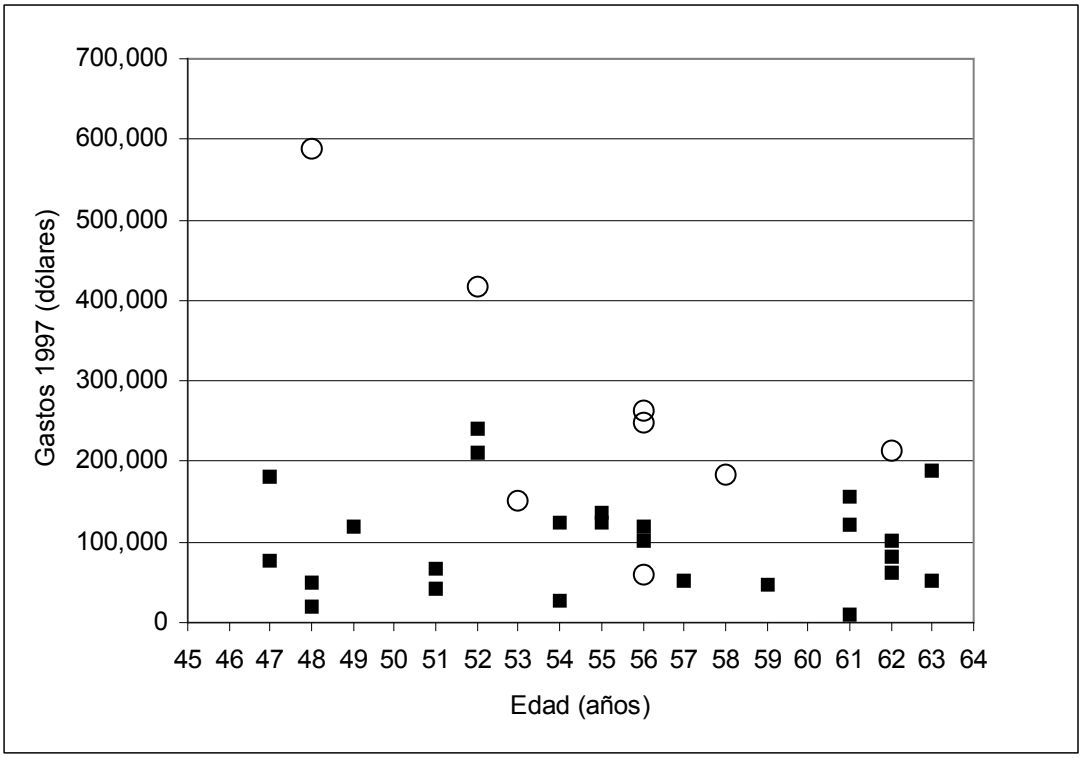

Fig. 1.1 Diagrama de dispersión para un conjunto de pacientes diabéticos.

A continuación se describen brevemente los términos empleados en esta definición.

Datos. Son un conjunto de hechos *F*, los cuales pueden ser ejemplares en una base de dato. En el ejemplo mostrado los datos son los valores de las dos variables que se tienen de una muestra de pacientes diabéticos.

Patrones. Se refiere a una expresión *E* en algún lenguaje *L* mediante la cual se describe hechos en un subconjunto F_E de *F*. *E* se denomina patrón si se puede expresar de una manera más general que una enumeración. Un ejemplo de patrón es la regla "Si los gastos de un paciente en 1997 son mayores a 180,000 dólares, entonces sus gastos en 1998 serán mayores a 125,000 dólares".

Proceso. Este término implica que hay varios pasos que involucran la preparación de datos, búsqueda de patrones, evaluación de conocimiento, así como refinamiento de la información. Estos pasos se pueden repetir en múltiples iteraciones.

No trivial. Significa que el proceso no puede ser totalmente automatizado. En el ejemplo mostrado puede ser útil obtener cálculos de promedios de edades, pero el proceso abarca más que el hecho de realizar cálculos, ya que involucra la búsqueda de estructura, modelos, patrones o parámetros en los datos.

Válidos. Significa que los patrones encontrados deben tener una buena predictibilidad, es decir, que se pueden aplicar los patrones en nuevos datos con un grado razonable de certeza. Si para predecir los pacientes más costosos de 1998 se usa la regla "Si la edad de un paciente es mayor a 55 años, entonces sus gastos en 1998 serán mayores a 125,000 dólares" se van a etiquetar equivocadamente muchos pacientes que en realidad

tuvieron gastos menores al umbral. Una regla más válida (aunque puede haber otras mejores) sería “Si los gastos de un paciente en 1997 son mayores a 180,000 dólares, entonces sus gastos en 1998 serán mayores a 125,000 dólares”.

Nuevos. Mediante la minería de datos se desea encontrar patrones no conocidos previamente por el usuario.

Potencialmente útiles. Cuando se buscan patrones se desea que éstos se puedan aplicar de alguna manera. Supóngase que la regla “Si la edad de un paciente es mayor a 55 años, entonces sus gastos en 1998 serán mayores a 125,000 dólares” es válida. Un ejemplo de utilidad de esta regla sería realizar algún tipo de medidas preventivas en pacientes menores a los 55 años.

Entendibles. Los patrones deben poder interpretarse y no ser solamente cajas negras para el usuario. En el ejemplo mostrado los patrones fueron reglas que son fáciles de interpretar.

La siguiente figura muestra el esquema iterativo e interactivo del KDD.

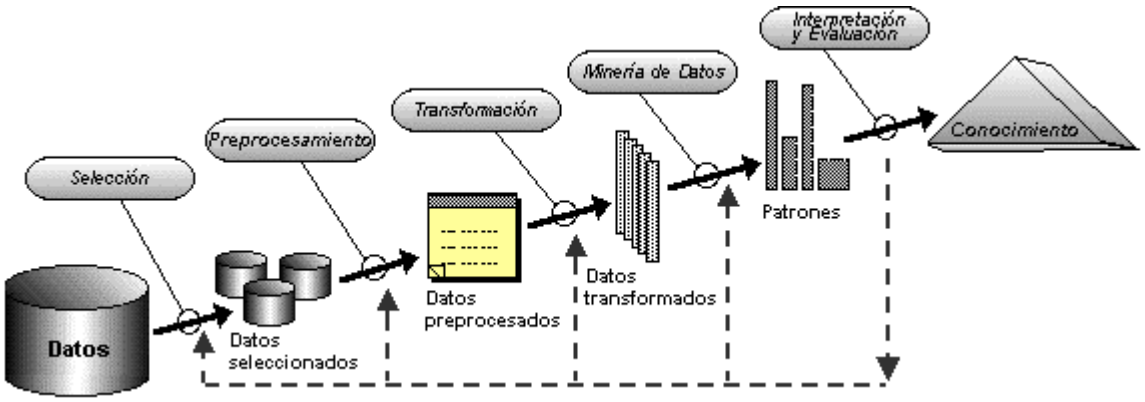

Fig. 1.2 Esquema iterativo e interactivo del KDD.

De acuerdo con Fayyad, Piatetsky-Shapiro y Smyth [11], los pasos involucrados en este esquema son los siguientes:

1. Aprender el dominio de la aplicación. Incluye tener el conocimiento *a priori* relevante así como los objetivos de la aplicación.
2. Crear un conjunto de datos objetivo. Incluye seleccionar un conjunto de datos o enfocarse en un subconjunto de variables o muestras de datos sobre los cuales se desea descubrir la información.
3. Limpieza de los datos y pre-procesamiento. Incluye operaciones básicas como eliminar el ruido y datos anómalos, decidir las estrategias para manejar los datos faltantes de campos, así como decidir tipos de datos y mapeos de valores desconocidos.
4. Reducción de datos y transformaciones. Incluye encontrar las características (también referidas como variables o atributos) útiles para representar los datos (dependiendo del objetivo de la tarea), reducir la dimensión de los datos, o transformar los datos.
5. Elegir la función de minería de datos. Incluye decidir la tarea a realizar por el algoritmo de minería de datos.

6. Elegir los algoritmos de minería de datos. Incluye seleccionar los métodos a usar para buscar los patrones en los datos, así como decidir qué modelos y parámetros pueden ser apropiados
7. Minería de datos. Incluye buscar los patrones de interés.
8. Interpretación. Incluye interpretar los patrones descubiertos y posiblemente regresar a uno de los pasos previos, así como la posible visualización de los patrones obtenidos, desechar patrones redundantes o irrelevantes y traducir los patrones útiles en términos que sean claros al usuario.
9. Usar el conocimiento descubierto: incluye incorporar el conocimiento en el desempeño del sistema, tomar acciones, o simplemente documentar y reportar a las partes interesadas, así como revisar y resolver conflictos potenciales con conocimientos previos.

Es importante resaltar que el proceso de KDD es iterativo e interactivo y por lo tanto se puede regresar a pasos anteriores del proceso en caso de ser necesario, como lo indican las flechas punteadas en la figura 1.2. Otro punto importante a mencionar es que en el esquema mostrado no se aprecia el enfoque multiperspectiva que se desea enfatizar en el presente trabajo. Sin embargo, el esquema presentado es importante porque en su momento sentó las bases para el desarrollo de la minería de datos.

1.3 Tareas de la minería de datos

En términos generales, la minería de datos se pueden agrupar en dos grandes tareas:

1. Descripción
2. Predicción

1.3.1 Descripción

El objetivo de la descripción, en ocasiones también referida como perfilación [27], es encontrar las características que definen a un grupo en particular. Generalmente este grupo específico es el conjunto de muestras que pertenece a una clase dada. De esta manera, lo que se busca con la perfilación es responder a la pregunta ¿por qué este grupo pertenece a la clase dada?

Existen diversas técnicas que nos ayudan a resolver esta pregunta. A continuación se mencionarán algunas de las técnicas más comunes.

Visualización

La visualización permite presentar los datos en distintas formas gráficas. Esta representación facilita la exploración simultánea de una gran cantidad de datos, de esta manera se puede obtener un panorama general de ellos y también es útil para detectar anomalías en los datos.

Las limitaciones para realizar visualizaciones se han reducido debido a los avances tecnológicos, aunque estos avances también implican que cada vez se dispone de más información a analizar. Aún así, las facilidades para visualizar los datos son mayores que años atrás.

En la minería de datos generalmente se tienen datos de un gran número de variables (decenas o centenas de variables y en algunos casos hasta miles de variables). Estos datos multivariados son difíciles de visualizar debido a nuestras limitaciones visuales. Existen algunas técnicas de visualización que permiten representar datos de más de tres dimensiones, sin embargo, todavía se limitan a un número reducido de dimensiones.

Algunas de las técnicas de visualización más comunes son: histogramas, gráficas de barras, gráficas de dispersión, matrices de gráficas de dispersión, curvas y superficies de nivel, gráficas de estrella, caras de Chernoff, curvas de Andrew, árboles y dendrogramas (diagramas de árbol jerárquicos). Se puede encontrar más información sobre estos métodos de visualización en [15] y [19].

Sumarización

La sumarización muestra los datos de una manera más reducida, permitiendo de esta manera calcular valores derivados de los datos originales registrados. Se puede considerar como una generalización de los datos y por lo mismo suele facilitar el aprendizaje ya que esto facilita el reconocimiento de patrones.

Otra ventaja de sumarizar los datos, es que pueden obtener nuevas variables que sean más significativas que las variables a detalle.

Una aplicación práctica de la sumarización es la comparación de clases, ya que al sumarizar los datos de distintas clases se puede realizar de manera rápida la comparación de ellas.

Al igual que la visualización, la sumarización también permite detectar anomalías en los datos.

Las medidas más comunes para sumarizar datos son: valores mínimos, valores máximos, media y desviación estándar. La ventaja de estas medidas es que son muy sencillas de calcular y proporcionan un panorama general de las variables.

Otras medidas que han resultado muy útiles para la descripción de los datos son las medidas ε y ε'. La definición y aplicación de estas dos medidas se verá con más detalle en capítulos posteriores.

Clustering

El *clustering* es la clasificación no supervisada de objetos (patrones) [5]. En una clasificación no supervisada no se dispone de un conocimiento previo sobre las clases de los datos y por lo tanto no se utilizan datos de entrenamiento. En lugar de esto se involucran algoritmos de agrupamiento.

Los algoritmos de agrupamiento clasifican los patrones, basándose en una medida de proximidad entre ellos, en grupos llamados *clusters*. Un *cluster* es un conjunto de entidades similares con base en una medida de similitud definida previamente. Esto implica que las entidades de *clusters* diferentes no son semejantes. Cuando se realiza el agrupamiento se espera que el grado de asociación en los objetos de la misma categoría sea grande y bajo entre los objetos de distintas categorías.

De acuerdo con Jain [17], los algoritmos de agrupamiento se pueden clasificar, de manera general, en dos tipos de técnicas: de *partición* y *jerárquicas*.

Una técnica de agrupamiento jerárquico impone una estructura jerárquica en los datos, la cual consiste en una secuencia anidada de particiones. Esta secuencia de particiones se crea a partir de una matriz de disimilitud. La estructura jerárquica resultante se representa mediante un diagrama llamado dendrograma. En las técnicas jerárquicas no se requiere conocer inicialmente el número de *clusters* de los datos ya que el dendrograma resultante permite decidir el número de *clusters* de los datos. Los algoritmos de esta técnica requieren muchos recursos de cómputo por lo que no son muy usados en conjuntos grandes de datos.

Una técnica de agrupamiento de partición crea una sola partición de los datos al aplicar una función criterio para obtener la mejor partición. Usualmente, estas técnicas operan en una matriz de patrones. A diferencia de las técnicas jerárquicas, en estas técnicas los patrones se pueden mover de un *cluster* a otro de manera que una partición inicial mala pueda ser corregida. Por lo general, en estas técnicas se requiere conocer de antemano el número de *clusters* en que se dividirán los datos.

El *clustering* es útil para explorar los datos ya que descubre relaciones entre los datos, agrupándolos de acuerdo a sus similitudes. También se puede emplear para realizar clasificación de los datos, ya que el usuario puede determinar la clase correspondiente a cada *cluster* resultante.

Reglas de asociación

Una regla de asociación es un enunciado probabilístico sobre la ocurrencia de ciertos eventos dentro de los datos. Tienen como objetivo identificar relaciones no explícitas entre las variables. Pueden ser de muchas formas, pero la más común es:

SI $A = x$ y $B = y$ ENTONCES C = z con probabilidad p.

Donde A, B y C son variables y p es la probabilidad condicional $p(C = z|A = x, B = y)$. La probabilidad $p(C = z|A = x, B = y)$ se refiere como la precisión o confianza de la regla y la probabilidad $p(A = x, B = y)$ se refiere como la cobertura o soporte de la regla. Se considera una buena regla aquella que tiene una precisión mayor, así como cobertura, a cierto umbral establecido.

La ventaja del uso de reglas de asociación es que son sencillas e interpretables.

La idea general de las reglas de asociación se originó del análisis de datos tipo “canasta de mercado” (*market-basket*), también conocido como “cesta de compra” [16], en donde se encuentran reglas de asociación que indican que si un cliente compró los productos x y y, entonces también compró el producto z con una probabilidad p. La aplicación directa a un rango amplio de problemas de negocio junto con la interpretabilidad de las reglas hicieron populares este método de minería de datos.

El algoritmo básico para encontrar reglas de asociación boolenas es el algoritmo *a priori*. [24]. A partir de este algoritmo surgieron variantes para mejorar su eficiencia: uso de tablas *hash* y de árboles para acceder a los datos, muestreo, selección de variables,

particiones. Una explicación detallada del algoritmo se puede encontrar en [16]. Dependiendo de los datos el algoritmo puede llegar a requerir muchos recursos computacionales.

1.3.2 Predicción

Mientras que en la descripción se busca responder la pregunta ¿por qué?, en la predicción lo que se busca es responder las preguntas:

- ¿Quién? Ejemplo: ¿quiénes son los mejores candidatos a ingresar en una universidad?
- ¿Cuál? Ejemplo: ¿cuál es la probabilidad de que un estudiante termine sus estudios?
- ¿Dónde? Ejemplo: ¿en que zona geográfica hay buenos prospectos de estudiantes?
- ¿Cuándo? Ejemplo: ¿cuándo es probable que un estudiante se gradúe?

Debido a la complejidad del mundo y a su ruido inherente, no se puede responder con certeza estas preguntas. Lo más conveniente es responder estas preguntas mediante probabilidades. Por ejemplo, para responder la pregunta ¿cuándo es probable que se gradúe un estudiante?, en lugar de dar una fecha específica, lo más conveniente sería obtener la distribución de la probabilidad de graduación de dicho estudiante con respecto al tiempo.

Búsqueda y predictibilidad

El objetivo de la predicción es establecer relaciones causales estadísticas entre las distintas variables para identificar patrones. Esta búsqueda de patrones puede verse como una búsqueda de predictibilidad, definiendo la predictibilidad como una medida del grado de reproducción de los patrones en los datos. Es decir, si en un conjunto de datos se encontraron ciertos patrones, la predictibilidad nos indica que tan probable es que esos patrones aparezcan en otro conjunto de datos estadísticamente similar. La predictibilidad varía como una función de las variables predictoras del problema y de los valores de estas variables. Por ejemplo, en un análisis inicial de los datos, se podría encontrar que la variable edad proporciona buena predictibilidad para determinar si un estudiante va a terminar sus estudios, y también se podría encontrar que la variable sexo no proporciona buena predictibilidad. Se pueden usar distintas medidas de predictibilidad. Sea $\boldsymbol{X}$ un vector de características n-dimensional que representa un ejemplar de un conjunto de datos con n variables; una medida de predictibilidad que es muy útil es $p(C|\boldsymbol{X})$, es decir, la probabilidad de que un vector dado $\boldsymbol{X}$ pertenezca a la clase C, sin embargo se pueden usar otras medidas.

La búsqueda de patrones o predictibilidad se realiza en espacios que dependen del tipo de datos del problema, como se aprecia en los siguientes ejemplos:

1. En la predicción de los clientes más probables de adquirir cierto producto, los patrones se pueden buscar en un espacio de variables sociodemográficas, donde cada variable sociodemográfica puede tener dos o más posibles valores. En este ejemplo un cliente se puede representar como un vector n-dimensional, donde n indica el número de variables sociodemográficas que se tienen.
2. En la clasificación de una imagen de satélite Landsat, los patrones se buscan en un espacio de siete variables, una variable por cada banda espectral de la imagen.

Cada variable puede tener 256 valores para indicar el nivel de gris en la imagen. En este ejemplo un pixel se puede representar como un vector de siete dimensiones.

3. En la búsqueda de texto dentro de un conjunto de documentos, los patrones se pueden buscar en un espacio de muchas dimensiones, donde las variables son las palabras más relevantes que se encuentran en el documento. Los valores de las variables pueden indicar el número de ocurrencias de la palabra en el documento. En este ejemplo, un documento se puede representar mediante un vector *n*-dimensional, donde *n* depende del número de palabras en los documentos.

Independientemente del tipo de datos y el tipo de valores de las variables, lo que tienen en común la mayoría de los problemas de minería de datos es que la búsqueda de patrones se realiza en un espacio muy grande. Considérese que en el primer ejemplo se tienen 100 variables sociodemográficas y que todas las variables son binarias. En este caso, se tienen 2^{100} posibles vectores de clientes. Y si nuestro conjunto de datos consiste solamente de 100,000 clientes, entonces la búsqueda de patrones se convierte en la búsqueda de una aguja en un pajar. Con el segundo ejemplo, ocurre lo mismo, ya que en total se pueden tener 2^{56} posibles vectores de pixeles. Si en el tercer ejemplo consideramos que un documento puede tener hasta 1,000 variables (una variable por cada palabra distinta) y que el número de una misma palabra por documento puede variar entre 0 y 49, entonces estamos hablando de un total de 50^{1000} posibles vectores, lo cual equivale aproximadamente a 2^{564} posibles valores. Si a la gran dimensión del espacio se añade que generalmente se dispone de muy pocos datos, entonces la búsqueda de patrones se complica. La cantidad de datos necesarios a menudo se incrementa exponencialmente si se desea mantener cierto nivel de precisión en las estimaciones realizadas. Este problema se conoce como la "maldición de la dimensionalidad" [15].

Para enfrentar el problema de "la maldición de la dimensionalidad" se pueden seguir dos enfoques:

1. Reducir la dimensión del espacio de búsqueda.
2. Realizar búsquedas inteligente en los datos.

Las dos maneras en que se puede reducir el espacio de búsqueda son: reducir el número de variables de los datos y reducir el número de posibles valores de las variables.

La manera más práctica para reducir el número de variables es mediante la selección de las variables más predictivas para nuestro problema específico. Esta selección se puede hacer después de realizar el análisis exploratorio y descriptivo de los datos. Otra manera de reducir el número de variables es aplicando alguna transformación con este fin, como puede ser el análisis de componentes principales. La desventaja de las transformaciones de este tipo es que dificultan la interpretación de los datos, lo cual es una gran desventaja en problemas de minería de datos. Sin embargo, este tipo de transformaciones es útil en problemas de clasificación de imágenes multiespectrales.

La reducción del número de posibles valores de una variable (cardinalidad de la variable) se puede hacer mediante la discretización de variables reales o realizando mapeos de rangos de variables discretas a un número más reducido de valores. Por ejemplo, en lugar de tener 100 posibles valores de la variable edad, donde cada valor indica los años de la persona, se pueden considerar sólo diez posibles nuevos valores, donde cada nuevo valor representa una década en lugar de un año. De esta manera el 1 correspondería a

una edad de 0 a 9 años, el 2 a una edad de 10 a 19 años, etc. A esta última técnica también se le denomina *coarse graining* [27].

Aún después de reducir la dimensión de los datos, es probable que la dimensión siga siendo grande, por lo que es conveniente buscar los patrones en los datos mediante alguna técnica de búsqueda inteligente, como son los algoritmos evolutivos, los cuales son útiles cuando se tiene un espacio muy grande y se cuenta con muy pocos datos.

En el capítulo 2 se detallará más en las distintas prácticas que se pueden emplear para afrontar "la maldición de la dimensionalidad".

Métodos de predicción

Existen varios métodos para predecir. Una manera de predecir sobre los datos es realizando una clasificación. Por ejemplo, si lo que queremos es predecir qué personas comprarán cierto producto, entonces se puede desarrollar un modelo mediante el cual se clasifique a las nuevas personas en "comprará el producto" y "no lo comprará". Hay una gran variedad de métodos para realizar clasificaciones. Un método de clasificación que en la práctica ha demostrado ser eficiente y que es sencillo de implementar es la clasificación Bayesiana, en particular la clasificación Bayesiana ingenua. Como ya se mencionó, el empleo de los algoritmos evolutivos también ha resultado ser muy útil, sobre todo cuando el espacio de búsqueda es muy grande y se dispone de pocos datos. En los últimos años también se ha incrementado el uso de redes neuronales artificiales, sin embargo, la desventaja de las redes neuronales es que dificultan interpretar de manera fácil los resultados de la clasificación.

Otra manera de predecir sobre los datos es empleando regresiones. Básicamente, en la regresión se busca predecir el valor numérico desconocido de una variable a partir de valores conocidos de otras variables.

En las siguientes secciones se presentará un panorama general de algunos de los métodos más populares para predecir sobre los datos.

Regresión

Mediante este método se establece una relación entre las variables de predicción (también conocidas como explicativas, independientes, o de regresión) para determinar el valor de la variable a predecir (denominada variable de respuesta, dependiente u objetivo). Por ejemplo, la predicción de los futuros costos de un determinado paciente a partir de su actual condición médica se puede hacer mediante un método de regresión.

Lo más común es que la variable de respuesta y las variables de predicción sean numéricas, pero también se pueden tener variables nominales. En el caso de que la variable de respuesta sea nominal, entonces el objetivo de la regresión es clasificar los datos.

Los modelos de la forma:

$$\hat{y} = a_0 + \sum_{i=1}^{p} a_i x_i \qquad \textbf{(1.1)}$$

donde:

- $\hat{y}$ es el estimado del valor real y
- a_i son los coeficientes de regresión
- x_i son las variables de predicción
- p es el número de variables de predicción

son conocidos como modelos de regresión lineal. En el caso más sencillo donde sólo se tiene una variable de predicción (regresión simple), el modelo genera una línea recta. En los casos donde hay más de una variable de predicción (regresión múltiple) se genera un plano de regresión. Los modelos de regresión lineal son muy usados debido a su simplicidad y a que generalmente tienen buen desempeño al predecir. Aún en los casos en que se sabe que la relación entre la variable de respuesta y las variables de predicción no es lineal. La razón de esto dada en [15] es que las relaciones lineales son una buena aproximación de relaciones no lineales.

En un problema real, lo más seguro es que el modelo no sea exacto, por lo que habrá un error entre el valor real de la variable a predecir y el valor estimado $\hat{y}$. Las diferencias entre los valores reales y los predichos se denominan residuos y se denotan por e:

$$y(j) = \hat{y}(j) + e(j) = a_0 + \sum_{i=1}^{p} a_i x_i(j) + e(j), 1 \le i \le n \qquad \textbf{(1.2)}$$

donde:

- n es el número de muestras del conjunto de datos
- j es la j-ésima muestra del conjunto de datos

La ecuación (1.2) se puede expresar en notación matricial como:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{e} \qquad \textbf{(1.3)}$$

donde:

- $\boldsymbol{y}$ es el vector de valores respuesta, de dimensión n x 1
- $\boldsymbol{a}$ es el vector de coeficientes de regresión, de dimensión (p + 1) x 1
- $\boldsymbol{e}$ es el vector de residuos, de dimensión n x 1
- $\boldsymbol{X}$ es una matriz, de dimensión n x (p + 1), compuesta por los vectores $\boldsymbol{X}_1$ ... $\boldsymbol{X}_p$ y el vector unitario, todos los vectores de dimensión n x 1

Para tener una predicción lo más precisa posible se deben encontrar los parámetros $\boldsymbol{a}$ que de alguna manera minimicen $\boldsymbol{e}$.

El método más popular para lograr esto es el método de mínimos cuadrados. Mediante este método se minimiza la suma de los errores al cuadrado, es decir, se minimiza:

$$\sum_{i=1}^{n} e(i)^2 = \sum_{i=1}^{n} \left(y(i) - \sum_{j=0}^{p} a_j x_j(i) \right)^2 \qquad \textbf{(1.4)}$$

donde:

p es el número de variables de predicción
$y(i)$ es el vector con los p valores reales observados en la i-ésima muestra de los datos
$x_j(i) = (1, x_1(i), ..., x_p(i))$ es el vector aumentado de la i-ésima muestra de los datos

Los coeficientes de regresión que minimizan (1.4) están dados por:

$$\boldsymbol{a} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\boldsymbol{y} \qquad (1.5)$$

Para resolver la ecuación (1.5) se requiere que la matriz ($\boldsymbol{X}^T\boldsymbol{X}$) sea invertible. Si el valor de n es pequeño o si las variables de predicción no son linealmente independientes, entonces lo más probable que la matriz ($\boldsymbol{X}^T\boldsymbol{X}$) no tenga inversa. Generalmente en los problemas de minería de datos se tienen suficientes muestras de los datos por lo que n no será pequeño. Lo que sí puede ocurrir más seguido es que las variables de predicción no sean linealmente independientes. En estos casos se debe hacer una selección de variables para eliminar la dependencia lineal en las variables. Si las variables de predicción no son totalmente linealmente independientes, pero casi lo son, entonces la solución encontrada (los coeficientes de regresión) puede ser inestable, lo que implica que al usar otras muestras de entrenamiento para obtener los parámetros, estos serán distintos. Este problema afecta cuando los coeficientes de regresión son el foco de interés (por ejemplo para determinar cuáles son las variables de predicción más importantes). Sin embargo, si lo que interesa más es la precisión de la predicción entonces el problema no afecta demasiado.

Clasificación

El problema de clasificación consiste básicamente en realizar una partición del espacio de características en regiones, una región para cada clase o categoría. [7]. Las clases se definen previamente y dependen del problema específico a resolver.

Un clasificador asigna un vector de características desconocido, es decir, sin categoría asignada, a una de las M clases de patrones. Las M clases se denotan como C_1, C_2, ... C_M y componen el conjunto $\boldsymbol{\Omega}$ de clases de patrones, $\boldsymbol{\Omega}$ = { C_1, C_2, ... C_M}. El clasificador divide el espacio de características en M regiones, llamadas *regiones de decisión*, denotadas $\Re_1$, $\Re_2$, ..., $\Re_M$.

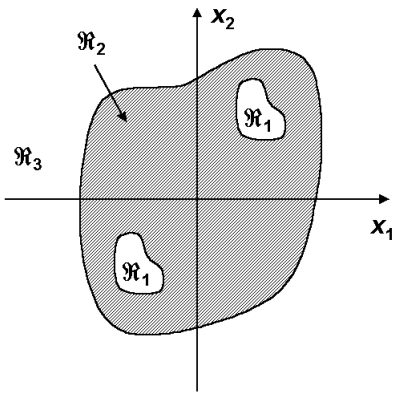

Fig. 1.3 Espacio de dos variables dividido en tres regiones de decisión.

La figura 1.3 muestra como se divide un espacio de patrones bidimensional en tres regiones de decisión. Se observa que los puntos del espacio de patrones están separados por superficies (las cuales son curvas en espacios bidimensionales), llamadas *superficies de decisión o límites de decisión*. Estas superficies de decisión dividen el espacio de características en M regiones de decisión.

Una manera para realizar la clasificación es modelando directamente las superficies de decisión. Sin embargo, en aplicaciones reales lo más probable es que las superficies de decisión no estén bien definidas por lo que las clases no pueden ser perfectamente separables.

Otra manera es usando funciones discriminantes para modelar las regiones de decisión. De esta forma un clasificador puede definirse implícitamente por un conjunto de M funciones. Sean $g_1(\boldsymbol{X})$, $g_2(\boldsymbol{X})$, ..., $g_M(\boldsymbol{X})$ funciones escalares, donde $\boldsymbol{X}$ representa un vector del espacio de características. Estas funciones, denominadas *funciones discriminantes,* son escogidas de tal forma que $g_i(\boldsymbol{X}) > g_j(\boldsymbol{X})$ para $i, j = 1, ..., M, j \neq i$. Es decir, para toda $\boldsymbol{X}$ en $\Re_i$, la i-ésima función discriminante tiene el valor más grande.

El problema central en el diseño de un clasificador consiste en especificar un conjunto apropiado de funciones discriminantes $g_1(\boldsymbol{X})$, $g_2(\boldsymbol{X})$, ..., $g_M(\boldsymbol{X})$ [23].

Generalmente las funciones discriminantes representan la probabilidad $p(C_k|\boldsymbol{X})$ de que el individuo $\boldsymbol{X}$ pertenezca a la clase C_k.

Existen muchos métodos para realizar clasificación. En las siguientes secciones se mostrará un panorama general de los métodos de clasificación Bayesiana, redes neuronales y algoritmos evolutivos.

Bayesiana

En la clasificación Bayesiana lo que se quiere determinar es la probabilidad de pertenencia $p(C_j|\boldsymbol{X})$ a una clase ω_i dado un vector $\boldsymbol{X}$. Esta probabilidad, también denominada como probabilidad *a posteriori*, se puede obtener usando el teorema de Bayes, el cual se expresa de la siguiente forma:

$$p(C_j \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid C_j)p(C_j)}{p(\boldsymbol{X})} \qquad \textbf{(1.6)}$$

donde:

$p(\boldsymbol{X} \mid C_j)$ es la probabilidad de que ocurra $\boldsymbol{X}$ dado que es un patrón perteneciente a la categoría C_j. Generalmente esta probabilidad es referida como verosimilidad (*likelihood*) de C_j con respecto a $\boldsymbol{X}$.

$p(C_j)$ es la probabilidad *a priori* de ocurrencia de la categoría C_j.

$p(\boldsymbol{X})$ es la probabilidad de que $\boldsymbol{X}$ ocurra independientemente de su categoría y está definida por:

$$p(\boldsymbol{X}) = \sum_{j=1}^{M} p(\boldsymbol{X} \mid C_j)p(C_j)$$

Las funciones discriminantes se pueden expresar en términos de las funciones de probabilidad $p(\boldsymbol{X}|C_i)$, $i = 1, ..., M$, y las probabilidades *a priori* $p(C_i)$, $i = 1, ..., M$ como:

$$g_i(\boldsymbol{X}) = p(\boldsymbol{X} \mid C_i)p(C_i) \text{ para } i = 1, ..., M \qquad \textbf{(1.7)}$$

Dado que la función logaritmo es una función monotónica creciente, generalmente es conveniente expresar las funciones discriminantes de la siguiente manera:

$$g_i(\boldsymbol{X}) = log\ p(\boldsymbol{X} \mid C_i) + log\ p(C_i) \text{ para } i = 1, ..., M \quad \textbf{(1.8)}$$

El clasificador que usa las funciones discriminantes definidas por (1.7) o (1.8) realiza la categorización mediante los siguientes pasos:

1. Se presenta el patrón ***X*** al clasificador;
2. Se calcula $g_i(\boldsymbol{X})$ para $i = 1, ..., M$;
3. El clasificador decide que:
   $\boldsymbol{X} \in C_k$ si $g_k(\boldsymbol{X}) > g_i(\boldsymbol{X})$ para toda $i \neq k$, $i = 1, ..., M$ **(1.9)**

El método de entrenamiento para el diseño de las funciones discriminantes consiste en los siguientes pasos:

1. Las funciones discriminantes se expresan en términos de los parámetros de $p(\boldsymbol{X}|C_i)$ y los parámetros $p(C_i)$.
2. Se estiman los valores de los parámetros de $p(\boldsymbol{X}|C_i)$ y los parámetros $p(C_i)$ a partir de un conjunto de datos de entrenamiento. Obsérvese que se pueden tener muchos valores de ***X*** dependiendo de la dimensionalidad del espacio.
3. Se asume que estos estimados son los valores verdaderos de los parámetros y se usan en las expresiones para las funciones discriminantes desarrolladas en el paso 1.

La clasificación Bayesiana es un método sencillo, pero tiene una alta complejidad computacional.

En [15] se puede encontrar más información sobre un tipo particular de la clasificación Bayesiana: la clasificación Bayesiana ingenua. Este tipo de clasificación reduce en gran medida la complejidad computacional y tiene un buen desempeño de predicción.

Redes Neuronales Artificiales

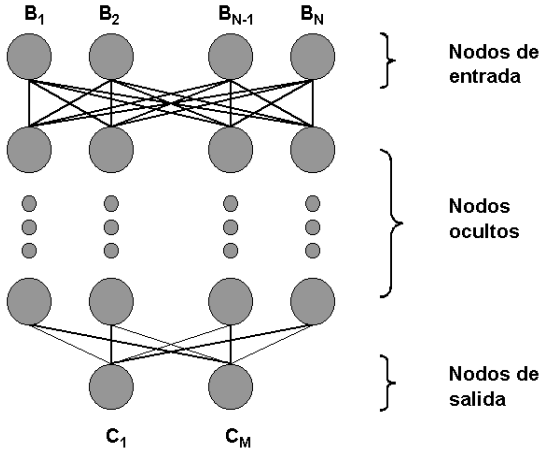

Fig. 1.4 Perceptrón multicapa

Una red neuronal artificial consiste en un número de nodos interconectados (equivalentes a las neuronas biológicas). Cada nodo es un *elemento simple de procesamiento*, que responde a entradas ponderadas que recibe de otros nodos [3].

Las entradas se ponderan mediante los pesos *w,* de esta manera, w_{ij} es el factor de ponderación entre un nodo *i* y un nodo *j*. La distribución de los nodos es referida como *arquitectura de la red*.

Existen muchos tipos de redes neuronales. Uno de los tipos más usados para la clasificación es el *perceptrón multicapa*, MLP (*Multi Layer*

Perceptron). En un perceptrón multicapa se tienen varias capas de nodos como se puede observar en la figura 1.4.

La primera capa es la *capa de entrada* y está compuesta por los nodos de entrada. En el caso de un MLP para la clasificación de datos *n*-dimensionales, el número de nodos de entrada es igual a *n*, el número de variables.

Después se tienen una o más capas intermedias, conocidas como *capas ocultas*. No existen reglas precisas para determinar el número de capas ocultas y nodos en estas capas. En términos generales, entre mayor sea el número de nodos en las capas ocultas, la red neuronal representará de mejor manera los datos de entrenamiento, pero a expensas de su capacidad de generalización. Lippman [21] propone una guía para la selección del número de capas ocultas y de nodos en estas capas.

La última capa es la *capa de salida*. El número de nodos de salida depende de como se usen las salidas para representar las clases. El método más simple es dejar que cada nodo denote una clase diferente.

En un MLP, las señales de entrada ingresan a la red mediante la capa de entrada. Posteriormente, van pasando a los nodos de las capas subsecuentes. Mientras la señal pasa de nodo a nodo se va modificando por los pesos asociados con la conexión. Cada nodo receptor, a excepción de los nodos de entrada, suma las señales ponderadas provenientes de todos los nodos de la capa anterior con los que está conectado.

El objetivo del entrenamiento de un MLP es ajustar los pesos w_{ji}, de manera que la red entrenada permita generalizar y predecir salidas a partir de entradas que no conoce. Para entrenar el MLP se alimenta la red con un patrón de entrenamiento. La salida obtenida se compara con la salida deseada y se calcula el error. Este error se va retro-propagando a través de la red y los pesos de las conexiones (usualmente los pesos iniciales se fijan de manera aleatoria) se van modificando. Este proceso de alimentación hacia adelante de las señales y la retro-propagación del error se repite en iteraciones hasta que se logra un error aceptable. Una vez que la red está entrenada se pueden clasificar patrones no conocidos por la red.

Los factores que afectan las capacidades de un MLP para clasificar son: el número de nodos, la arquitectura de la red, el tamaño del conjunto de entrenamiento y el tiempo de entrenamiento, es decir, las iteraciones durante el entrenamiento.

Las principales ventajas de usar un MLP son:

- Las clases no tienen que ajustarse a un tipo de distribución;
- Permiten realizar los cálculos en paralelo;
- Pueden simular regiones de decisión complejas.

Sin embargo, los MLPs requieren disponer de una muestra de datos de entrenamiento representativa, ya que son más sensibles a los datos de entrenamiento que los métodos estadísticos. La principal desventaja de los MLPs es que requieren mucho procesamiento numérico en la etapa de entrenamiento, por lo que puede consumir mucho tiempo entrenar un conjunto grande de datos. Otra desventaja es que no hay reglas precisas para determinar la arquitectura de la red neuronal, ni para determinar los parámetros

empleados en la regla delta generalizada. Además las redes neuronales dificultan una fácil interpretación de los resultados.

Para profundizar más sobre las redes neuronales se puede consultar [22] y [12]. La primera referencia es una introducción a las redes neuronales y la segunda se enfoca en los algoritmos, aplicaciones y técnicas de programación de las redes neuronales.

Algoritmos evolutivos

Así como las redes neuronales artificiales se basan en el comportamiento de las redes neuronales biológicas, los algoritmos evolutivos se basan en la teoría de selección natural originada por Darwin.

En los algoritmos evolutivos se tiene una población de individuos, la cual va evolucionando a lo largo de varias generaciones. Cada individuo de la población representa una solución candidata a un problema dado. Para medir la calidad de la solución de un individuo se usa una función de aptitud (*fitness*). La evolución de los individuos hacia mejores individuos se realiza a través de procesos selectivos basados en la selección natural (es decir mediante la supervivencia y reproducción del más “fuerte”) y usando operadores genéticos de entrecruzamiento y mutación. Mientras mejor sea la calidad de un individuo, entonces será más probable que su “material genético” pase a los individuos de generaciones posteriores.

Los algoritmos evolutivos son métodos muy flexibles de búsqueda y se pueden aplicar a muchos tipos de problemas. Dependiendo del problema se elige la representación de los individuos y la función de aptitud de los individuos.

En problemas de minería de datos los algoritmos evolutivos se pueden usar para descubrir reglas de predicción. En este caso, los individuos corresponden a una regla de predicción, o a un conjunto de reglas de predicción. La evaluación de estas reglas se hace mediante la función de aptitud. La ventaja de las reglas de predicción es que son muy fáciles de interpretar.

Los algoritmos evolutivos son útiles en la búsqueda de patrones en espacios muy grandes. Su utilidad en este tipo de problemas se debe a que realizan una búsqueda global en lugar de realizar búsquedas locales. Esta búsqueda global se logra mediante el operador de entrecruzamiento, así como trabajando simultáneamente con un conjunto de individuos, es decir con un conjunto de soluciones, en lugar de una sola solución.

Los algoritmos genéticos son un paradigma de algoritmos evolutivos que enfatizan el entrecruzamiento como el operador principal de búsqueda exploratoria [15]. En los algoritmos genéticos iniciales los individuos se representaban mediante cadenas binarias, pero actualmente existen otro tipo de representaciones como cadenas de enteros o de valores reales. En [13] y [14] se puede profundizar sobre los algoritmos genéticos y evolutivos.

A lo largo de la sección 1.3 se han mostrado los métodos o técnicas más comunes para las principales tareas de la minería de datos. A continuación se presentan dos tablas en donde se muestra un resumen de los métodos más comunes.

En la tabla 1.1 se muestran los tipos de datos que pueden usarse en los métodos más comunes de la minería de datos.

Tipo de datos	Descripción			Predicción			
	Visualización y Sumarización	Reglas de asociación	*Clustering*	Regresión	Clasificación Bayesiana	Redes Neuronales	Algoritmos Evolutivos
Numéricos Continuos	X		X	X	X	X	X
Numéricos Discretos	X	X	X	X	X	X	X
Nominales sin orden	X	X	X		X	X	X
Nominales con orden	X	X	X	X	X	X	X

Tabla 1.1 Tipos de datos que pueden usarse en los métodos más comunes de minería de datos.

La tabla 1.2 muestra un resumen con las principales características de las técnicas más comunes que hay en la minería de datos.

Técnica de minería de datos	Principales características
Visualización	Facilita la exploración simultánea de una gran cantidad de datos. Permite obtener un panorama general de los datos. Util para detectar anomalías en los datos.
Sumarización	Generaliza los datos. El uso de variables sumarizadas en ocasiones facilita el reconocimiento de los patrones. Facilita la comparación de distintas clases de interés. Es útil para detectar anomalías en los datos.
Clustering	Descubre relaciones entre los datos. Además de describir los datos permite realizar clasificaciones. Requiere muchos recursos computacionales.
Reglas de asociación	Descubre relaciones entre los datos. Permite resolver problemas tipo “canasta de mercado”. Dependiendo de los datos puede llegar a requerir muchos recursos computacionales.
Regresión Lineal	Es un método sencillo. Se asume independencia lineal de las variables. Generalmente tienen buen desempeño. No se puede usar cuando se dispone de muy pocos datos.
Clasificación Bayesiana	Es un método sencillo. Tiene alta complejidad computacional. Su variante ingenua reduce la complejidad computacional y generalmente tiene un buen desempeño de predicción, aún cuando las variables no sean independientes

Técnica de minería de datos	Principales características
Redes Neuronales Artificiales	Es un método complejo. Las clases no tienen que ajustarse a un tipo de distribución. Permiten realizar cálculos en paralelo. Pueden simular regiones de decisión complejas. Requieren disponer de una muestra de datos representativa. Se requiere experiencia para determinar la arquitectura y parámetros de la red neuronal. No es fácil interpretar los resultados. Requieren mucho procesamiento numérico en la etapa de entrenamiento.
Algoritmos Evolutivos	Son métodos flexibles de búsqueda. Son útiles en la búsqueda de patrones en espacios muy grandes. Se pueden aplicar a muchos problemas. Además de la predicción permiten la descripción de los datos. Requieren mucho procesamiento numérico en la etapa de entrenamiento.

Tabla 1.2 Características de las principales técnicas de minería de datos.

1.4 Minería de datos en el mundo real.

La minería de datos es un proceso que se puede aplicar en un gran rango de problemas, que pueden ir desde problemas académicos a aplicaciones reales.

Los objetivos de un problema académico son distintos a los objetivos de un problema de la vida real: generalmente en un problema académico se desea estudiar algún modelo de minería de datos y compararlo contra otros modelos, mientras que en una aplicación real lo que se busca es resolver un problema particular, generalmente complejo. Debido a esto, en una aplicación real no basta con resolver el mismo problema usando distintos métodos y comparar los resultados de estos métodos. Esto es útil para evaluar el desempeño de los distintos métodos, pero también hay otros factores importantes a considerar, como el tiempo necesario para entrenar o correr los modelos, así como la facilidad de interpretación y el grado de acción que permiten los resultados de los modelos.

En la sección 1.2 se presentó el proceso a seguir para obtener conocimiento de los datos. En este proceso hay varios puntos que son de vital importancia para resolver un problema de minería de datos:

1. Antes que nada hay que entender y plantear correctamente el problema que se requiere resolver. Como ya se mencionó, los problemas reales de minería de datos generalmente son complejos, y por lo mismo requieren descomponerse en varios problemas. Es importante que las metas del problema planteado sean realistas y no estén más allá de los recursos con los que se cuentan.
2. Posteriormente hay que analizar los datos disponibles y entender sus limitaciones. Si durante este análisis se descubre que los datos no son adecuados para resolver el problema, o que se necesitan más datos, o datos de otro tipo, entonces lo más apropiado es colectar más datos. Sin embargo, no siempre es posible obtener más datos, por lo que es importante entender las limitaciones de los datos disponibles.

3. Con base en los dos puntos anteriores se debe plantear la solución y métodos más adecuados para resolver el problema. Dependiendo del problema quizás se requiera más de una solución y método.

Las siguientes secciones tratarán sobre los tres puntos planteados.

1.4.1 La naturaleza multiperspectiva del problema

Los problemas del mundo real son muy complejos y generalmente su resolución implica el cumplir con varios objetivos. Para resolver un problema del mundo real primero se deben establecer los objetivos y sus prioridades.

Considérese el ejemplo de un banco que desea realizar minería de datos para detectar comportamientos fraudulentos en transacciones de datos, por ejemplo para detectar el uso fraudulento de tarjetas de crédito. En este caso se podría realizar una clasificación de las transacciones de los clientes, en dos clases: fraudulenta o no fraudulenta. Sin embargo, para el banco podría ser más útil tener una estimación de la probabilidad de fraude y con base en esta estimación seleccionar los casos más probables de fraude, para lo cual se puede aplicar una regresión logística. Esto requiere contar previamente con un conjunto de datos de casos que se sepa son fraudulentos, sin embargo en aplicaciones reales no siempre se puede contar con esta información. Otro factor a considerar es si se desea estimar la detección de fraudes considerando el tiempo, lo cual es razonable dado que los fraudes regularmente son eventos que una persona realiza recurrentemente a lo largo del tiempo. También se puede considerar realizar análisis espacial ya que para el banco podría ser útil modelar patrones geográficos de actividad fraudulenta. Otro factor a considerar es el tipo de acciones que se realizarán cuando se presuma que existe un fraude, ya que el tipo de medidas podría ahuyentar a buenos clientes clasificados equivocadamente como fraudulentos, además estos clientes podrían iniciar una campaña de queja contra el banco y repercutir negativamente en la imagen del banco.

Este ejemplo muestra la complejidad de la minería de datos en problemas del mundo real y también muestra como un mismo problema puede tener más de una solución. Debido a esta naturaleza multiperspectiva de los problemas reales se deben definir bien los objetivos, así como sus prioridades.

1.4.2 Los datos y sus limitaciones

Es importante tener en cuenta que no se puede realizar una buena minería de datos si la calidad de los datos es mala: pocos datos, datos con mucho ruido, errores en los datos, etc. Aún cuando hay técnicas para aminorar algunos de estos problemas, si los datos sólo tienen información basura, por más que se busque no se encontrará información valiosa. Sin embargo, si se dispone de material razonable, entonces será más factible encontrar conocimiento dentro de los datos. Encontrar oro en la minería de datos significa encontrar predictibilidad, es decir, encontrar patrones que siguen siendo válidos en conjuntos diferentes (aunque estadísticamente similares).

Datos adquiridos con propósito distinto a la minería de datos

Frecuentemente la minería de datos se realiza sobre un conjunto de datos que se obtuvo para un propósito diferente a obtener conocimiento de los datos. Por ejemplo, una

universidad puede tener una base de datos en donde registre el historial de sus estudiantes con el objetivo de realizar trámites escolares, y hasta después de varios años usar esta base de datos para realizar minería de datos, quizás para investigar el alto grado de deserción de estudiantes. Este tipo de ejemplo puede ser muy común debido a que los avances tecnológicos han incrementado el número de bases de datos y cada vez hay más personas que desean aprovechar la información que tienen disponible.

En los casos en que se requiere realizar minería de datos en un conjunto ya existente de datos, se puede tener la dificultad de que los datos simplemente no son predictivos con respecto a lo que se desea predecir. Es decir, que para los fines deseados, los datos no son de valor. Sin embargo, ¿como saber que en realidad no se puede extraer patrones valiosos de los datos? Quizás no se está buscando correctamente, o quizás no se está usando la herramienta adecuada. Existen varios aspectos que pueden afectar la predictibilidad que hay en los datos, como se mencionará a continuación.

Conocimiento específico del dominio

Para realizar una adecuada minería de datos es preciso incorporar conocimiento específico del dominio en el proceso de recolección de los datos. Por ejemplo, si se desea realizar minería de datos en una base de datos de estudiantes de una universidad para determinar quienes son los mejores estudiantes, para empezar, el personal de la universidad debería definir en qué consiste ser un buen estudiante y de acuerdo con esta definición identificar los campos de su base de datos que definen a un buen estudiante. En caso de que se necesite mayor información a la disponible en la base de datos, el personal de la universidad tendrá una mejor intuición que alguien externo a la compañía, para pensar en que otros datos son valiosos para recolectar.

Datos espaciales y temporales no homogéneos

Otro elemento importante que degrada la predictibilidad presente en los datos es la heterogeneidad espacial o temporal en la distribución estadística que soporta los datos. Es decir, que lo más conveniente es tener conjuntos de datos estadísticamente similares aún cuando los conjuntos corresponden a distintos periodos de tiempo o distintas zonas geográficas. Para saber si dos conjuntos son estadísticamente similares se pueden emplear medidas estadísticas sencillas como son la media y la varianza de las variables.

La tendencia y dispersión de los datos son dos factores que afectan la predictibilidad de los datos. Ambos factores pueden aparecer en la muestra de la cual se obtuvieron los datos, así como en el contenido de los datos adquiridos por cualquier instrumento. Generalmente la tendencia en los datos está asociada con el uso de muestras estadísticamente diferentes, como pueden ser las muestras adquiridas en distintos periodos de tiempo, o la combinación de datos adquiridos de distintas fuentes. Por ejemplo, si se combinan encuestas de clientes, adquiridas dentro de una tienda, con encuestas de no-clientes, adquiridas en un centro comercial, se puede encontrar que las características sociodemográficas de los dos tipos de muestras son completamente diferentes. Este requerimiento de homogeneidad estadística también aplica al conjunto de datos que se usa para entrenar los modelos empleados y el conjunto de datos en el cual se desea predecir. Por muy bueno que sea el modelo de minería de datos planteado, no se obtendrán buenos resultados si el conjunto en el cual se desea predecir es estadísticamente muy diferente al conjunto usado para entrenar el modelo.

La importancia de las escalas espaciales y temporales

Otro elemento importante que afecta la predictibilidad se asocia con determinar las escales espaciales y temporales más convenientes para un problema dado. Por ejemplo, si una universidad desea realizar una campaña a nivel masivo para atraer estudiantes de todo el país, entonces, el uso de información sociodemográfica a nivel de municipios, puede ser el más adecuado, pero si la campaña se desea realizar a nivel local, digamos para cierta zona de una ciudad, entonces se requiere un mayor detalle de los datos, como podría ser por código postal, o incluso a nivel individual. Lo mismo sucede con la escala de tiempo, si se desean predecir los estudiantes que se graduarán en un periodo de años, entonces quizás lo más conveniente sea tener datos con una escala de meses, en lugar de días. Las escalas espaciales y temporales dependen del problema específico de minería de datos.

Incertidumbre y barreras de predictibilidad

La incertidumbre en la predictibilidad se puede deber a diversos factores: errores en los datos, tamaño limitado de la muestra, mala elección de las variables, así como cambios en la distribución probabilística de los datos debido a la heterogeneidad espacial y temporal de los dato.

Existen varios tipos de anomalías que pueden tener los datos, como son: valores faltantes, valores extremos (fuera de rango) o errores en los valores (por ejemplo que se tenga un valor alfanumérico en lugar de uno numérico). Algunas acciones para reducir estas anomalías son: ignorar los valores anómalos, eliminar la variable que presenta errores, eliminar el registro que tiene errores o reemplazar los valores. Cada una de estas acciones tiene sus ventajas y desventajas, así como repercusiones dependiendo del método de minería de datos que se vaya a emplear y de las características de los datos. Por ejemplo, los árboles de decisión son robustos a datos faltantes y las redes neuronales son sensibles a los valores extremos, aunque por otra parte son robustas al ruido extremo y a las variables no significativas.

El problema del tamaño limitado de la muestra puede afrontarse mediante la reducción de dimensionalidad o mediante métodos inteligentes de búsqueda. En principio, estos dos problemas no presentan una barrera fundamental sobre cuanta predictibilidad puede obtenerse de los datos. Sin embargo, la heterogeneidad espacial y temporal de los datos son problemas más sutiles e importantes en la barrera de la predictibilidad.

En la predicción se desea establecer una relación entre una variable objetivo y un conjunto de variables de los datos. Como ya se ha mencionado, este conjunto de variables se obtiene de un espacio de variables muy grande (si tenemos n variables, binaria entonces existen 2^n posibles conjuntos de variables). Diferentes opciones de variables estarán asociadas con diferentes grados de predictibilidad, y cada conjunto tendrá una barrera de predictibilidad que no se podrá rebasar. Desafortunadamente, no es posible conocer con exactitud el límite exacto de predictibilidad, sin embargo, después de analizar los datos se puede tener una intuición de cuáles son las barreras de predictibilidad. Aún cuando se puedan aminorar los problemas que causan la barrera de predictibilidad siempre existirá una barrera ya que para poder eliminarla completamente se necesitarían tener todas las variables que predicen a la variable objetivo, así como conocer sus relaciones, y esto es prácticamente imposible.

Si se está realizando por primera vez minería de datos en un problema dado, entonces se pueden realizar mejoras significativas de la predictibilidad. Si el problema no es nuevo también se pueden lograr mejoras sustanciales si se encuentran nuevos indicadores que sean más predictivos, o si se cambia a un mejor modelo. Sin embargo, si el modelo ya se encuentra en la barrera de predictibilidad, entonces sólo se lograrán mejoras muy pequeñas por más esfuerzos que se realicen para mejorar la predictibilidad. Un indicador se refiere a la combinación de una o más variables, y posiblemente uno o más valores de ellas, los cuales están asociados a la pertenencia a alguna clase de interés.

1.4.3 Planteamiento de la solución y selección del método

En el ejemplo de detección de fraudes en transacciones de clientes se mostró la complejidad de un problema del mundo real.

Dada la naturaleza multiperspectiva de los problemas del mundo real, lo más probable es que se plantee más de una solución al problema. La mayoría de los artículos y libros no discuten los puntos prácticos para resolver un problema de minería de datos, sin embargo, existen puntos generales que se deben considerar al resolver un problema de este tipo [20]:

- Lo mejor es resolver el problema con métodos sencillos e ir cambiando a métodos más complejos sólo en casos necesarios. Por ejemplo, en el caso de la detección de fraudes, el uso de series de tiempo y modelado espacial pueden ser útiles en un principio, pero quizás no valga la pena usar estos modelos si un método más sencillo proporciona una precisión adecuada.
- Para resolver una tarea de minería de datos a menudo se necesita transformar o abstraer la definición original del problema real. Hay que tener cuidado de no alejarse demasiado del problema original, sino solamente descartar los detalles que son irrelevantes al problema. Hacer esto es un arte que requiere práctica.
- En la solución de un problema de minería de datos se involucran dos partes: el minero que extrae la información y el experto del dominio. Para seleccionar apropiadamente una tarea de minería de datos es necesario que tanto el minero como el experto tengan un entendimiento completo del problema, así como de la tarea seleccionada. Generalmente esto requiere invertir tiempo del minero para entender el dominio y tiempo del experto del dominio para entender los principios generales de la tarea de minería de datos seleccionada.
- Una recomendación conocida en la práctica de la estadística aplicada es que es mejor aproximar la solución del problema correcto que resolver exactamente el problema incorrecto. En ocasiones los académicos tienden a preferir algoritmos complejos en lugar de probar con métodos menos complejos que proporcionan una solución más sencilla, robusta e interpretable.

Una vez que se planteó como se va a resolver el problema hay que seleccionar el método más adecuado para implementar la solución.

Un enfoque para seleccionar un método de minería de datos es comparar el desempeño de un conjunto fijo de algoritmos en diversos conjuntos de datos. Otro enfoque similar es, dado un problema, aplicar todos los algoritmos disponibles para poder seleccionar el que mejor resuelve el problema, lo cual claramente es ineficiente. Además, la definición de "mejor" depende de las metas específicas de la aplicación, así como de los datos.

La selección de un método apropiado debe tomar en cuenta las restricciones de la aplicación, así como los datos requeridos por los métodos. Por ejemplo, en una aplicación de medicina, a menudo es importante un entendimiento completo del conocimiento antes de usarlo, mientras que el tiempo requerido para obtener un resultado suele ser menos importante que la calidad deseada del resultado.

Generalmente, la preselección de algoritmos se basa en los requerimientos de la aplicación. Esta preselección de métodos forma la base para la selección posterior de un solo método que cumple con todos los requerimientos y restricciones de la aplicación. Existe una variedad de propiedades que componen las restricciones de la aplicación [20]:

- Restricciones de tiempo de corrida. ¿Es importante el tiempo de ejecución de un algoritmo dado? Quizás en una tarea de clasificación el tiempo no sea un factor importante al realizar el entrenamiento, pero sí lo sea para clasificar nuevos datos.
- Tipo del modelo. Existen diferentes modelos que pueden realizar tareas similares. La elección de un modelo depende de los requerimientos exactos del usuario. Las propiedades principales de los modelos son:
    - Su habilidad para generalizar.
    - Su interpretabilidad.
    - Su compactación.
- La aptitud (*fitness*) del modelo. Por ejemplo, la exactitud y tasa de error al predecir en nuevos datos. La complejidad de un modelo también se debe tomar en cuenta al definir la aptitud de un modelo para resolver una aplicación particular.
- Restricciones del sistema. Estos incluyen el tipo de hardware disponible así como el acceso a herramientas específicas de KDD.

Capítulo 2: Predictibilidad y enfoque multiperspectiva

2.1 Paisaje de predictibilidad

Como ya se mencionó en el capítulo 1, la minería de datos está ligada a la búsqueda de predictibilidad. Existen dos consideraciones importantes en esta búsqueda:

1. Determinar las características del espacio sobre el cual se realiza la búsqueda
2. Definir la medida de predictibilidad a usar. Con base en esta medida, lo que se busca es encontrar las variables y los valores de las variables que son predictivos.

Las medidas de predictibilidad se pueden ver como funciones de altura en el espacio de los vectores de características. De esta manera, se forma una topografía a la cual se le denomina paisaje de predictibilidad [27].

En la figura 2.1 se muestra un ejemplo de paisaje de predictibilidad de dos variables métricas. Las variables métricas son aquellas cuyos valores guardan cierto orden entre sí, por ejemplo, la edad es una variable métrica. La primera variable de la figura representa la contribución de los individuos que tienen seguro contra daño a terceros. La segunda variable representa las contribuciones de los individuos que tienen seguro contra incendios. Entre mayor sean los valores de estas variables, mayores son las contribuciones. La clase que se quiere predecir es la compra de una póliza para casas rodantes. En la gráfica se observa que mientras mayores sean los valores de las dos variables, aumenta la probabilidad de la clase.

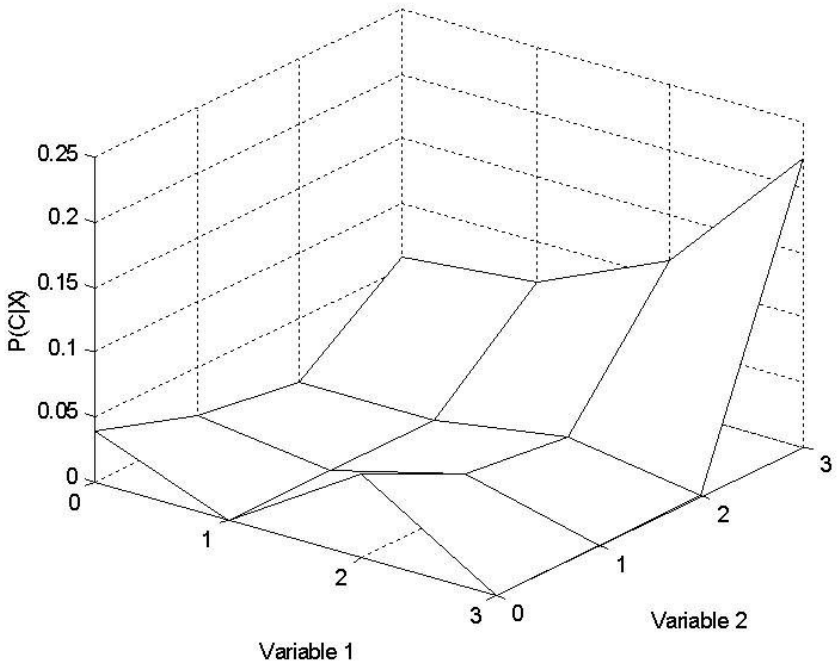

Fig. 2.1 Ejemplo de un paisaje de predictibilidad de dos variables métricas. La medida de predictibilidad usada es *p*(*C*|*X*).

El paisaje de predictibilidad de la figura 2.1 es relativamente sencillo. Incluso se puede modelar linealmente. Sin embargo, considérese la figura 2.2. En este caso se observa que el paisaje de predictibilidad es más complejo y multimodal.

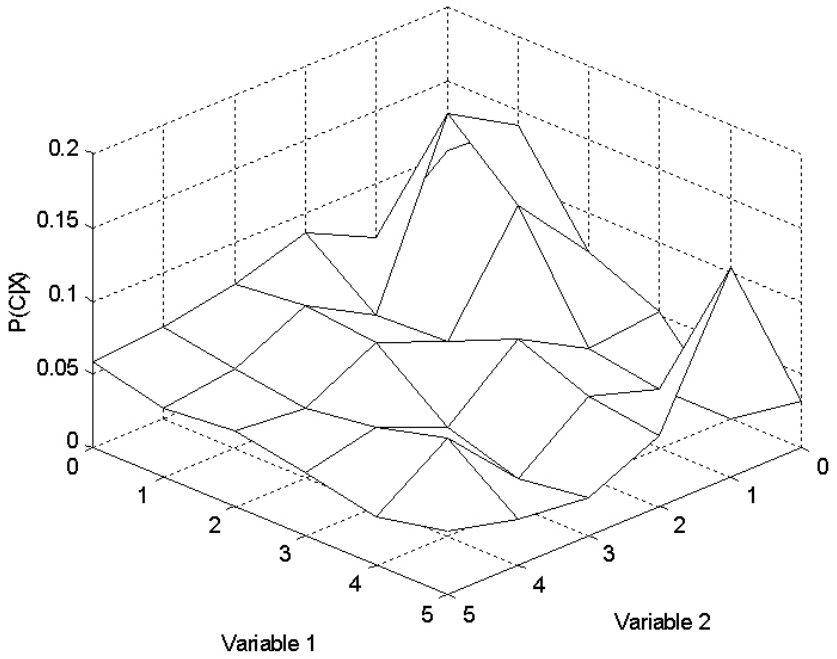

Fig. 2.2 Ejemplo de un paisaje de predictibilidad de dos variables métricas. La medida de predictibilidad usada es *p*(*C*|*X***). En este caso, el espacio es multimodal.**

El ejemplo corresponde al paisaje de predictibilidad de dos variables sociodemográficas. La primera variable representa el porcentaje de solteros en determinada zona geográfica y la segunda variable representa el porcentaje de personas sin religión, también por zona geográfica. La clase a predecir es la misma que en el primer ejemplo.

En este caso, no se puede aplicar una regresión lineal multivariada, o alguna otra función sencilla para obtener una buena aproximación, porque las relación entre la probabilidad *p*(*C*|***X***) y las dos variables es multimodal. Si hubiera suficientes datos para tener por lo menos cierto grado de confidencia, se podría dejar que los datos "hablaran por sí mismos", es decir que se podría establecer la relación entre la variable de salida y las variables de predicción con base en relaciones observadas entre estas variables en el conjunto de entrenamiento. Esto significa que se requieren varias observaciones para cada posible combinación de las variables de entrada. Sin embargo, si el espacio de características es muy grande y la muestra de datos es pequeña entonces este enfoque no es viable. Desafortunadamente, esta situación es muy común en problemas reales de minería de datos. Esto implica que en muchos problemas del mundo real, no se van a tener observaciones para la mayoría de las combinaciones de las variables. En el capítulo 1 se presentaron, de manera general, varias formas de afrontar este problema. En este capítulo se ahondará más en las soluciones presentadas.

2.2 La maldición de la dimensionalidad

Los problemas típicos de la minería de datos están asociados con cientos, o incluso miles de variables y cada una de estas variables puede tener más de un valor. El número de combinaciones de los valores-variables puede ser muy grande, incluso para problemas relativamente pequeños.

Sea n el número de variables que involucra el problema de minería de datos y sea a_i el número de posibles valores de la variable i, donde $1 \leq i \leq n$. Entonces el número m de combinaciones de los valores de las características está dado por:

$$m = \prod_{i=1}^{n} a_i \qquad (2.1)$$

Si se tienen cien variables con diez posibles valores, entonces $m = 10^{100}$. Para tener una idea de qué tan grande es este número: 10^{100} es mayor que el número de átomos en el universo observable, el cual es aproximadamente 10^{80}.

El ejemplo anterior no es un caso extremo, incluso es un caso muy conservador ya que existen problemas de minería de datos con miles de variables. Dado que la gran dimensionalidad es un factor presente en la mayoría de los problemas de minería de datos, entonces se requieren métodos efectivos para afrontar la gran dimensionalidad. Estos métodos consisten en:

1. Reducir la dimensión del espacio de búsqueda.
2. Realizar búsquedas inteligente en los datos.

2.3 Reducción del espacio de búsqueda

Considérese que se desean predecir los costos y de un paciente con base en 40 variables binarias $\boldsymbol{X} = [\, x_1, \ldots, x_{40} \,]$ que indican su condición médica (por ejemplo, cada variable indicando la presencia o ausencia de una enfermedad o síntoma). Es decir, que se desea obtener $y(\boldsymbol{X})$. Se puede proponer un modelo lineal de la forma:

$$y(\boldsymbol{X}) = a_0 + \sum_{i}^{40} a_i x_i \qquad (2.2)$$

En este modelo se requieren encontrar 41 parámetros a_i, $0 \leq i \leq 40$, lo cual aparentemente es sencillo. Sin embargo, aunque el modelo es simple, se está haciendo el supuesto de que las relaciones entre las variables de predicción son lineales. Si se asume que las relaciones entre y y $\boldsymbol{X}$ son deterministas, es decir, que sólo se tiene un valor de y por cada posible $\boldsymbol{X}$, y se ajusta el modelo para eliminar el error, entonces se necesitarían 2^{40} muestras de distintos vectores $\boldsymbol{X}$ para poder hacer una estimación relativamente precisa de y para cada posible valor de $\boldsymbol{X}$. Es prácticamente imposible tener tal cantidad de datos. Este escenario se complica si se considera que, en la mayoría de los casos, la relación entre y y $\boldsymbol{X}$ no es determinista. Esto significa que se tiene una distribución de y para cada vector distinto $\boldsymbol{X}$. En este caso, no sólo se necesita una muestra de cada vector $\boldsymbol{X}$, sino varias muestras para poder estimar la distribución de y. Aunque este problema se puede resolver de manera más sencilla mediante una regresión lineal, su planteamiento es útil

para comprender la relevancia que tiene el tamaño del espacio de búsqueda o características en la resolución de un problema de minería de datos.

Como ya se mencionó en el capítulo 1, existen varias maneras de reducir la dimensionalidad del espacio de búsqueda, las cuales consisten básicamente reducir el número de variables, reducir el número de valores de las variables y realizar transformaciones que ayuden a reducir la dimensionalidad. En las siguientes secciones se detallarán estos métodos.

2.3.1 Coarse graining (granulado grueso)

El *coarse graining* consiste en reducir los valores de una variable a un conjunto más pequeño de valores. En el caso de que se tenga una variable con valores reales que se cambien a valores enteros, a este proceso se le conoce como discretización. La diferencia con la discretización es que el *coarse graining* no se aplica solamente para cambiar valores reales a discretos, también se aplica para reducir el número de valores enteros de una variable. También se puede hacer *coarse graining* a nivel de variables, es decir reduciendo el número de variables como se verá en la siguiente sección.

La manera más elemental para hacer *coarse graining* consiste en hacer un *binning* de las variables. Supóngase que una de las variables que se tiene de una encuesta de 100 participantes es la edad y que la encuesta se hizo a personas entre 18 y 80 años. Esto significa que se pueden tener hasta 63 posibles valores de edad. Si la distribución de la edad de los participantes fuera uniforme, entonces para cada posible valor de edad se tendrían entre 1 y 2 participantes y por lo tanto la inferencia estadística basada en este número pequeño de muestras por edad se dificulta. Para aminorar este problema se pueden agrupar las edades en rangos, esto es posible porque no hay muchas diferencias significativas entre personas de 63 años y personas de 64 años (por mencionar un ejemplo), sin embargo, sí es posible que haya diferencias significativas entre personas de 18 años y personas de 30 años.

Una manera práctica de realizar el *coarse graining* consiste en dividir el rango total de la variable en un número reducido de rangos que sean del mismo tamaño. Por ejemplo, si la variable representa un porcentaje, entonces se pueden usar rangos para representar cada decena del porcentaje. Sin embargo, hay que tomar en cuenta que esta división no es la más conveniente para todas la variables, ya que los rangos adecuados para hacer *coarse graining* se deben seleccionar dependiendo de la distribución de la variable y de las características del problema. Por ejemplo, si la encuesta del ejemplo de arriba se aplica en una universidad, tanto a alumnos como maestros, lo más probable es que la mayoría de los participantes tenga un rango de edad entre 18 y 23 años. En este caso lo más conveniente sería no perder el detalle en ese rango de edad y usar rangos de edad más amplios para el resto de las edades.

Cuando se hace el *coarse graining* hay que tener cuidado de no perder predictibilidad. Usar muy pocos rangos podría implicar que se pierde detalle y predictibilidad. Por el contrario, usar muchos rangos podría implicar que se tienen predicciones no confiables estadísticamente, es decir, muy pocas muestras por rango. Esto último ocurre aún en el caso de que se tengan suficientes datos ya que si el número de variables es grande, el espacio de búsqueda sigue siendo muy grande a pesar de la reducción del número de valores.

Por ejemplo, considérese el caso de la predicción de los costos con base en las 40 variables binarias. Aún cuando el número de valores por variable es reducido (sólo 2 valores), el espacio de búsqueda sigue siendo muy grande ya que se tienen 2^{40} posibles valores de $\boldsymbol{X}$.

En estos casos se puede hacer una suposición sobre un modelo que se espera que se ajuste a los datos. Existen muchos tipos de modelos estadísticos, cada uno con distinto grado de simplicidad o complejidad, así como supuestos sobre los cuales el modelo es más útil. Los modelos pueden ir desde una regresión lineal hasta una red neuronal.

También se puede dejar que los datos "hablen por sí mismos". En este caso no necesariamente se propone un modelo, sino que se construye una estimación de la relación probabilística entre las variables de los datos. Anteriormente ya se mostraron los inconvenientes de este enfoque, así que de alguna manera se debe poder reducir más el espacio de búsqueda o realizar una búsqueda inteligente en el espacio como se verá en otra sección más adelante.

Como ya se mencionó al principio de la sección, otra manera de hacer *coarse graining* es al remover por completo algunas variables. Dependiendo del problema de minería de datos, habrá algunas variables que se puedan descartar fácilmente, pero también habrá otras variables para las cuales no sea obvio si se deben descartar o no. Para esto se necesitará una medida que nos indique la utilidad o predictibilidad de la variable. Esto se verá con más detalle en la siguiente sección.

Para ahondar más sobre el *coarse graining* se introducirá cierta notación que extiende los posibles valores asociados con una variable dada. Sea el símbolo "*"otro posible valor de la variable y que indica que la variable puede tomar cualquiera de sus posibles valores. Por ejemplo, considérese la probabilidad condicional $p(C|x_1x_2)$, la cual indica la probabilidad de pertenecer a la clase C dado que se tienen valores específicos de las variables x_1 y x_2, donde las variables representan el estado de salud de una persona. Por simplicidad se asume que estas dos variables son binarias y que representan algún síntoma de la persona. En este caso, se tiene 1 si el síntoma está presente y 0 si no lo está. Entonces, $p(C|11) = p(C|x_1 = 1, x_2 = 1)$ representa la probabilidad de que la persona pertenece a la clase C dado que presenta los dos síntomas. Usando la nueva notación, $x_i = *$ significa que la variable x_i puede tomar cualquiera de los valores 1 o 0. De esta manera $p(C|1*) = p(C|x_1 = 1, x_2 = *) = p(C|x_1 = 1, x_2 = 0) + p(C|x_1 = 1, x_2 = 1) = p(C|10) + p(C|11)$ es la probabilidad de tener el síntoma 1, independientemente del síntoma 2. La probabilidad $p(C|1*)$ es la probabilidad marginal de la clase C dada $x_1 = 1$. Un individuo siempre se asocia con un conjunto completo especificado de valores (podría haber casos en que no se tengan todos los valores, pero para efectos del ejemplo esos casos no se están considerando) y en ocasiones más de un individuo se puede asociar con el mismo conjunto de valores. Al incluir el símbolo "*" se pueden asociar vectores de características con grupos de individuos. Por ejemplo, considerando el ejemplo anterior de dos variables, el vector de características $\boldsymbol{X} = [\, x_1 = 1, x_2 = 1 \,]$ representa a los individuos que tienen dichos valores particulares para las dos variables, mientras que $\boldsymbol{X} = [\, x_1 = 1, x_2 = * \,]$ representa a los individuos que tienen $x_1 = 1$ y *cualquier* valor en la variable x_2. Obviamente, los vectores que involucran el símbolo "*" agrupan a más individuos que los que no lo involucran.

2.3.2 Selección de variables mediante las funciones *ε* y *ε'*

La selección de las variables a usar en un problema de minería de datos depende en gran medida de la perfilación o predicción que se está realizando. Para el caso de la perfilación o predicción de la pertenencia a una clase, existen dos funciones que son muy útiles para determinar qué tan importantes son las variables o valores de las variables de predicción. Estas dos medidas son *ε* para valores de las variables y *ε'* para las variables. La definición de estas dos medidas es la siguiente [27]:

$$\varepsilon(C \mid \boldsymbol{X}) = \frac{N_X\left[p(C \mid \boldsymbol{X}) - p(C)\right]}{\sqrt{N_X p(C)\left[1 - p(C)\right]}} \qquad \textbf{(2.3)}$$

donde:

- $N_{\boldsymbol{X}}$ es el número de observaciones asociadas con el vector de características $\boldsymbol{X}$.
- $p(C|\boldsymbol{X})$ es la probabilidad de que una observación asociada con el vector de características $\boldsymbol{X}$ pertenezca a la clase C. Se calcula como el número de observaciones asociadas con el vector de características $\boldsymbol{X}$ y que pertenecen a la clase C, entre el número de observaciones asociadas con el vector de características $\boldsymbol{X}$.
- $p(C)$ es la probabilidad de que una observación pertenezca a la clase C. Se calcula como el número de muestras que pertenecen a la clase C entre el número total de muestras.

La medida *ε* puede ser usada para variables discretas o continuas, así como para variables métricas o no métricas. Como se mencionó al principio del capítulo, las variables métricas son aquellas cuyos valores guardan cierto orden entre sí. *ε* mide la confiabilidad estadística de la diferencia de un conjunto de observaciones versus un conjunto obtenido de una distribución aleatoria, que es la hipótesis nula. Por ejemplo, para $\boldsymbol{X}$ = 1*, *ε* mide el grado en que el valor x_1 = 1 se asocia con la clase, independientemente del valor de la variable x_2. Entre mayor sea el valor de *ε*, mayor será el grado de confidencia con el que uno puede desechar la hipótesis nula de que $\boldsymbol{X}$ no es predictiva de la clase. En general, los valores de *ε* > 2 indican que el valor de la variable, o la combinación de valores de variables, es predictiva de la clase. De manera similar, para $\boldsymbol{X}$ = *0, *ε* mide el grado en el que el valor x_2 = 0 se asocia con la clase independientemente del valor de la variable x_1.

Además de usar *ε* para determinar los valores de las variables más importantes de acuerdo a que tan predictivas son éstas, también puede usarse como medida de altura en el paisaje de predictibilidad. En el ejemplo de la predicción de costos a partir de 40 variables de condición médica, donde las variables representan enfermedades o síntomas, $\varepsilon(x_1=1,\ x_2=*,x_{40}=*)$ es una medida de la predictibilidad asociada con la clase de membresía (por ejemplo la clase de pacientes con costos más bajos) para los individuos que tienen la enfermedad o síntoma x_1 (el valor 1 indica presencia y el valor 0 indica ausencia), independientemente del resto de sus enfermedades o síntomas.

$$\varepsilon'(x_i) = \frac{\langle x_i \rangle_C - \langle x_i \rangle_{\sim C}}{\sqrt{\frac{\sigma^2_{iC}}{N_{iC}} + \frac{\sigma^2_{i\sim C}}{N_{i\sim C}}}} \qquad \textbf{(2.4)}$$

donde:

- $\langle x_i \rangle_C$ es la media de la variable x_i sobre todas las observaciones que pertenecen a la clase C.
- $\langle x_i \rangle_{\sim C}$ es la media de la variable x_i sobre todas las observaciones que pertenecen a una clase distinta a C (generalmente la clase complementaria de C).
- σ^2_{iC} es la varianza de la variable x_i sobre todas las observaciones que pertenecen a la clase C.
- $\sigma^2_{i\sim C}$ es la varianza de la variable x_i sobre todas las observaciones que pertenecen a una clase distinta a C (generalmente la clase complementaria de C).
- N_{iC} es el número de observaciones de la variable x_i que pertenecen a la clase C.
- $N_{i\sim C}$ es el número de observaciones de la variable x_i que pertenecen a una clase distinta a C (generalmente la clase complementaria de C).

La medida ε' puede ser usada para variables continuas o discretas, pero a diferencia de ε sólo se puede usar para variables métricas. La restricción de métrica de las variables se ve claramente de la ecuación (2.4), ya que la variable x_i debe ser métrica para que la diferencia de las medias tenga sentido. Las variables binarias también se pueden considerar como métricas. De esta manera, en el cálculo de ε' también se pueden considerar variables cuyos valores no representan alguna medida que se puede ordenar (como la edad), sino alguna pertenencia a alguna de dos categorías, por ejemplo, si se es hombre o mujer.

El papel de ε' es determinar qué variables son predictivas de la clase de pertenencia. Si $(\langle x_i \rangle_C - \langle x_i \rangle_{\sim C})$ es significativamente diferente de cero y positivo, entonces los valores altos de x_i son predictivos de la clase. Y si es significativamente diferente de cero y negativo, entonces los valores bajos de x_i son predictivos de la clase. De manera similar que ε, los valores $\varepsilon' > 2$ típicamente indican que la variable correspondiente es predictiva de la clase de pertenencia. Dado que ε' involucra más datos que ε, se puede considerar que ε' tiene más confianza estadística que ε.

Las variables que tienen valores bajos de ε' se pueden desechar porque no son muy predictivas. Sin embargo, se debe tener cuidado al remover variables, ya que la variable aislada podría ser poco predictiva, pero en combinación con otra variable pudiera ser que fuera muy predictiva.

Es importante hacer notar que las medidas ε y ε' no son únicas y que no necesariamente son las mejores medidas. Sin embargo, en la práctica han resultado ser muy útiles porque miden adecuadamente si las variables y sus valores son predictivos, además son fáciles de calcular.

Mediante las medidas ε y ε', u otras medidas semejantes, se pueden determinar los indicadores más importantes para la perfilación y predicción. Como ya se ha visto en algunos ejemplos, ε se puede calcular para variables solas o para combinaciones de variables. Por ejemplo, si las variables para determinar los costos médicos de una persona son *edad* e *ingresos*, entonces sería de esperar que ε(*edad* = adulto, *ingresos* = bajos) fuera mayor que ε(*edad* = joven, *ingresos* = altos), mostrando de esta manera la

interacción entre las variables *edad* e *ingresos* para determinar los costos médicos de una persona. La desventaja de calcular ε para combinaciones de variables es que si son muchas variables se pueden llegar a tener miles de combinaciones, lo que trae como consecuencia que se tengan pocas observaciones para un vector dado $\boldsymbol{X}$ y por ende poca confiabilidad estadística. Debido a esto, no se aconseja calcular ε para combinaciones de muchas variables. Para aminorar este problema se pueden hacer combinaciones solamente de las variables más importantes, las cuales se pueden elegir mediante ε' en caso de ser variables métricas.

2.3.3 Transformación de variables

Una de las transformaciones más útiles para la reducción de características, es el análisis de componentes principales.

El análisis de componentes principales es una técnica estadística que consiste en transformar los datos linealmente para eliminar su correlación [18], es por esto que es una herramienta muy útil para analizar datos que tienen una alta correlación. Dicha técnica también es útil para analizar conjuntos de datos de distintos tipos. Se ha usado durante muchos años y su base matemática está bien documentada.

Los datos altamente correlacionados son datos en los cuales los valores de una variable pueden usarse para predecir los valores correspondientes de otra variable. Esta redundancia no está presente en los datos no correlacionados. Por lo tanto, los datos correlacionados pueden representarse de manera más compacta por un conjunto de variables no correlacionadas.

Esta técnica se basa en la varianza y covarianza de un conjunto de datos. La varianza es una medida de dispersión de una variable del conjunto de datos. La covarianza es una medida de dispersión entre dos variables. En el análisis de componentes principales se asume que la varianza de los datos se puede usar como una medida del contenido de información de los datos.

El análisis de componentes principales de las variables-*x* originales determina una transformación lineal que convierte toda la variabilidad de los datos originales en nuevas variables-*y*. La transformación es determinada de tal manera que la primera variable-*y* contiene la cantidad más grande posible de la varianza total. La segunda variable-*y* contiene la cantidad más grande posible de la varianza restante, así sucesivamente.

La pequeña cantidad de varianza contenida en las últimas variables a menudo es muy pequeña, por lo que se puede despreciar. Cuando estas variables son desechadas se reduce la dimensionalidad de los datos y la varianza total de los datos transformados es menor que la varianza total de los datos originales debido a la cantidad de varianza contenida en las variables desechadas. El nuevo conjunto de *k* variables, donde *k* es menor a la dimensión original, tiene mayor contenido de información que el que tienen *k* variables originales.

Las variables-*y* transformadas que resultaron del análisis de componentes principales no están correlacionadas. Generalmente, el número de variables-*y* transformadas resulta menor mientras mayor sea la correlación en la variables-*x* originales.

Una de las desventajas del análisis de componentes principales es que al realizar la transformación se pierde la facilidad de interpretación de las variables, es decir que las variables resultantes no están asociadas a un significado directo como en el caso de las variables originales.

2.4 Búsqueda inteligente

Lo más probable es que aún después de reducir la dimensionalidad del espacio, éste siga siendo muy grande. En estos casos lo mejor es emplear una búsqueda inteligente de los patrones y predictibilidad.

Se requiere realizar una búsqueda inteligente, ya que una enumeración aleatoria de todas las posibles combinaciones de valores de las variables es imposible en la práctica. Existen varias heurísticas para realizar búsqueda en espacios con alta dimensionalidad, los cuales varían desde técnicas sencillas de búsquedas locales, como *hill-climbing*, a técnicas más sofisticadas, como los algoritmos evolutivos o el recocido simulado.

Cada método tiene sus ventajas y desventajas, así como un rango de problemas donde son más eficientes. Esto significa que un método que es bueno para un tipo de problemas podría no serlo para otro tipo de problemas. La eficiencia de algunos métodos puede depender significativamente de sus parámetros. Es por esto, que se recomienda que si no se tiene una buena experiencia con algún método en particular, mejor se opte por un método más sencillo y transparente.

Al emplear un método se debe tener un balance entre la exploración y la explotación del espacio de búsqueda. Demasiada exploración significa que se está invirtiendo mucho tiempo en buscar patrones y ya no se encuentran patrones más predictivos de los que se tienen. Por otra parte, demasiada explotación significa que se está buscando dentro de un conjunto limitado de características siendo que existen más combinaciones de variables predictivas en otras partes del espacio de búsqueda.

2.4.1 Algoritmos genéticos

En el capítulo anterior se presentó un panorama general sobre las ventajas de los algoritmos evolutivos. En esta sección se hablará un poco de los métodos más tradicionales de búsqueda y al final se enfatizará cuales son las ventajas de los algoritmos genéticos sobre los métodos tradicionales de búsqueda.

En la literatura sobre búsqueda y optimización se mencionan tres principales tipos de métodos de búsqueda [13]:

1. Basados en cálculo
2. Enumerativos
3. Aleatorios

Métodos basados en cálculo

Estos métodos se subdividen en dos clases principales:

1. Indirectos.
2. Directos.

Los métodos indirectos buscan los extremos locales solucionando un conjunto, usualmente no lineal, de ecuaciones que resultan de establecer el gradiente de la función objetivo igual a cero. Esta es una generalización multidimensional de la noción elemental de cálculo de puntos extremos, como los mostrados en la figura 2.3. Dada una función suave, no restringida, para encontrar un posible pico se restringe la búsqueda a aquellos puntos que tengan pendiente cero en todas las direcciones.

Los métodos directos buscan óptimos locales basándose en la función y moviéndose en una dirección relacionada con el gradiente local. Esta es la noción de *hill-climbing*: para encontrar el mejor punto local se escala la función en la dirección más pronunciada permitida.

A pesar de que ambos tipos de métodos se han mejorado y extendido, presentan las siguientes limitaciones que no los hacen adecuados para todo tipo de problemas:

- Ambos métodos tienen alcances locales. Por ejemplo, supóngase que se tiene un espacio como el de la figura 2.3. Si se empieza la búsqueda en el punto marcado con una cruz, entonces no se encontrará el pico más alto. Una vez que se alcanza el pico local, las mejoras subsecuentes deben ser buscadas a través de un reinicio aleatorio u otra serie de técnicas.
- Dependen de la existencia de derivadas. Aún si se permiten aproximaciones numéricas de las derivadas, sigue siendo una desventaja severa.

En los problemas de minería de datos del mundo real es raro contar con las condiciones ideales requeridas para este tipo de búsqueda. Regularmente los espacios de características de los problemas de minería de datos están llenos de ruido, discontinuidades (recuérdese que se cuenta con muy pocos datos que no cubren todo el espacio) y presentan una topografía muy rugosa. Debido a esto, no se pueden aplicar los métodos basados en el cálculo para realizar la búsqueda de patrones.

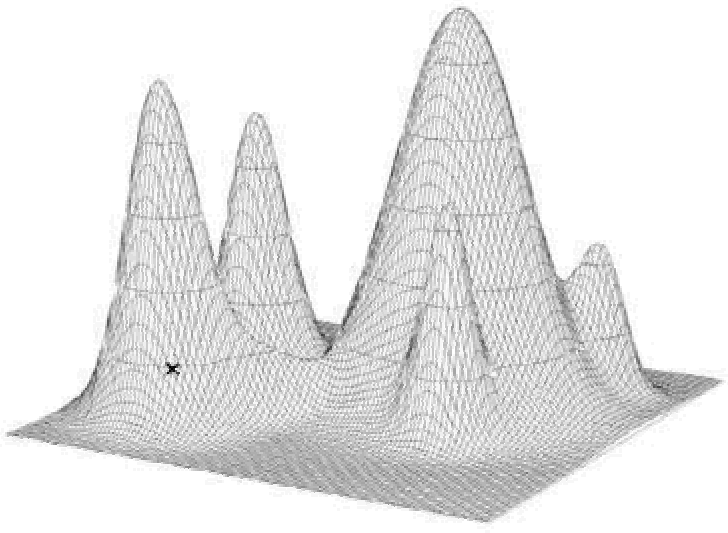

Fig. 2.3 Ejemplo de un espacio de características de dos variables. El espacio es multimodal.

Métodos enumerativos

Existen muchas variantes de este tipo de métodos. La idea principal de estos métodos consiste en dejar que los datos hablen por sí mismos, es decir fijarse en los valores de la función objetivo para cada punto del espacio de búsqueda, uno a la vez. La idea es sencilla y por lo mismo atractiva, pero en las secciones anteriores ya se han presentado ejemplos donde se concluye que este enfoque es imposible de realizar debido a que el espacio de búsqueda es inmensamente grande. Aún usando variantes como la programación dinámica, la búsqueda sigue restringida a espacios de tamaño moderado. Debido a esto, los enfoques enumerativos no son convenientes para resolver problemas de minería de datos del mundo real.

Métodos aleatorios

Estos métodos han surgido para contrarrestar las desventajas de las técnicas basadas en cálculo y enumerativas. En este tipo de métodos se realiza una exploración aleatoria que busca y guarda los mejores resultados. Sin embargo, a largo plazo, no se puede esperar que sean mejores que los métodos enumerativos. Es importante diferenciar entre búsquedas estrictamente aleatorias y técnicas "randomizadas". En estas últimas no necesariamente se realizan búsquedas completamente al azar, más bien se realizan elecciones aleatorias para guiar una búsqueda dirigida. Tal es el caso de los algoritmos genéticos y del recocido simulado. El punto a resaltar es que las técnicas de búsqueda randomizadas no tienen que implicar que la búsqueda se realiza sin dirección.

Eficiencia de los métodos convencionales de búsqueda

Las características mencionadas de los métodos tradicionales de optimización llevan a concluir que no son lo suficientemente robustos, o al menos no lo son para la mayoría de las aplicaciones del mundo real. Esto por supuesto, no implica que no sean útiles. Los esquemas mencionados, y sus combinaciones híbridas, han sido usados exitosamente en muchas aplicaciones, sin embargo, cuando se quiere resolver problemas más complejos, se requiere de otros métodos.

Para tener un panorama más claro de los métodos mencionados, considérese la figura 2.4. Esta figura muestra, a lo largo del eje *x*, un espectro de problemas, y a lo largo del eje *y* la eficiencia que tienen los distintos esquemas para resolver el espectro de problemas. En la gráfica se muestran tres curvas, una curva para cada esquema. La curva del esquema especializado indica un gran desempeño para problemas unimodales, pero mucha ineficiencia en el resto de los problemas. La curva del esquema enumerativo señala un desempeño ineficiente en los distintos problemas. Por último, a pesar de que la curva del esquema robusto indica que no hay un desempeño óptimo para un tipo específico de problemas, se observa un comportamiento eficiente en todo el espectro de problemas. Se puede mejorar la eficiencia de un esquema robusto combinándolo con algún método específico para resolver algún problema en particular.

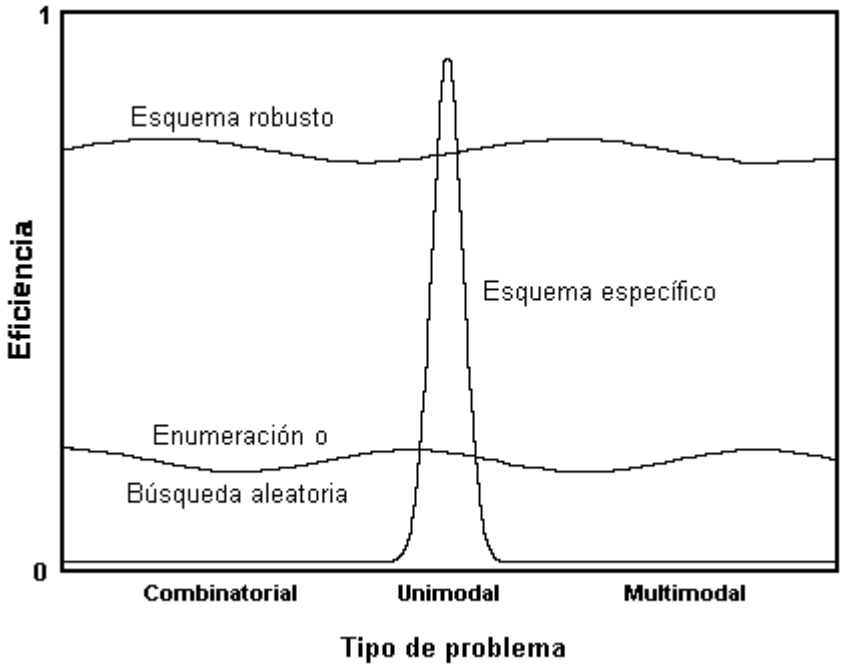

Fig. 2.4 Muchos esquemas tradicionales trabajan bien en un dominio de problema reducido. Los esquemas enumerativos y exploraciones aleatorias trabajan de manera ineficiente en espectro amplio de problemas. Un método robusto trabaja bien en un espectro amplio de problemas.

Metas de la optimización

La teoría de optimización comprende el estudio cuantitativo de los óptimos, ya sea de una función o proceso, así como los métodos para encontrarlos [14].

En esta definición hay que resaltar que se diferencia entre los óptimos buscados y los métodos para encontrarlos.

Cuando se resuelven problemas de optimización, generalmente se le da importancia a encontrar los óptimos y no a los procesos para encontrarlos, es decir que el enfoque es “lo importante es la meta, no el camino”. Sin embargo, en problemas de minería de datos del mundo real no se puede desestimar el camino. Como ya se ha mencionado, los problemas de minería de datos del mundo real tienen múltiples objetivos. Lo más probable es que el desempeño en tiempo sea un objetivo importante.

Supóngase que se quiere implementar un sistema para la bolsa de valores que prediga en tiempo real la probabilidad de que una acción suba, o baje, de precio. En este ejemplo, no se puede implementar un sistema que tenga muy buena precisión pero cuyo algoritmo tarde minutos en encontrar la mejor respuesta. Así como en este ejemplo el tiempo fue un factor importante, pueden existir otros factores de mayor relevancia que la precisión obtenida.

Para dar otro ejemplo, considérese a un hombre de negocios que debe tomar decisiones. Generalmente se juzgan sus decisiones cuando hace una selección adecuada dentro del tiempo y recursos disponibles. Nunca se juzga por una obtención del mejor criterio. Lo

más probable es que la convergencia a un óptimo no sea la cuestión más importante en los negocios. La obtención del óptimo es menos importante en sistemas complejos, como son los problemas de minería de datos del mundo real.

Ventajas de los algoritmos genéticos

Los algoritmos genéticos superan la robustez de los métodos tradicionales al diferir en cuatro aspectos fundamentales:

1. *Los algoritmos genéticos trabajan con una codificación del conjunto de parámetros, no con los parámetros en sí*. La ventaja de la codificación de los parámetros es que permite resolver problemas de manera muy general y por lo mismo no se tienen las restricciones de los métodos específicos (continuidad, existencia de derivadas, unimodalidad, independencia de variables, etc)
2. *Los algoritmos genéticos buscan una población de puntos, no un solo punto*. El peligro de buscar óptimos a partir de un solo punto es que se encuentren óptimos locales y no globales. Al realizar la búsqueda partiendo de varios puntos se incrementan las posibilidades de encontrar mejores óptimos locales e incluso encontrar los óptimos globales.
3. *Los algoritmos genéticos usan funciones objetivo, no derivadas u otros conocimientos auxiliares*. A diferencia de métodos específicos que requieren el uso de información auxiliar, los algoritmos genéticos sólo requieren el uso de una función de aptitud que mida la calidad de las soluciones que se van encontrando. Esta característica permite resolver un espectro amplio de problemas, pero esta generalidad también implica que el desempeño del algoritmo genético muy probablemente no será mejor que un método específico para resolver un problema dado. Sin embargo, existen maneras de incorporar información auxiliar, relativa al problema, para mejorar su desempeño.
4. *Los algoritmos genéticos usan reglas de transición probabilísticas, no reglas deterministas*. El uso de elecciones aleatorias es una herramienta para guiar la búsqueda en otras regiones del espacio de búsqueda con probabilidad de mejorar los óptimos encontrados al momento. No se trata simplemente de tomar decisiones aleatorias, sino de guiarlas mediante el azar.

2.5 Modelos de minería de datos en el paisaje de predictibilidad

En las secciones anteriores se ha mencionado que se puede ver la minería de datos como la búsqueda de predictibilidad. En esta sección se mostrará como se relacionan los modelos de minería de datos y el paisaje de predictibilidad. Existen muchos modelos de minería de datos y el alcance de esta tesis no comprende hacer un análisis exhaustivo de diferentes clases de modelos, por lo que en esta sección se enfocará solamente la atención en un modelo que en la práctica han resultado ser muy útil, el modelo de clasificación Bayesiana ingenua.

2.5.1 Modelos de minería de datos como plantillas topográficas en el paisaje de predictibilidad

Como ya se ha mencionado, la minería de datos consiste en descubrir patrones en los datos. Los patrones son relaciones $F(\boldsymbol{X})$ entre las variables de entrada, $\boldsymbol{X}$, y la variable de salida, F. Cualquier relación $F(\boldsymbol{X})$ se puede graficar dentro del paisaje de predictibilidad, como una función de altura. Generalmente, lo que interesa encontrar en la topografía son

los picos del paisaje, sobre todo si la variable es una probabilidad. También se pueden usar otras funciones, por ejemplo si se desea encontrar la relación de las variables que minimizan costos, entonces en lugar de una función de probabilidad se puede usar una función de costos. Además de los picos, en ocasiones también será de interés encontrar las regiones en donde hay un incremento brusco en la altura, esto es útil para determinar las variables discriminatorias de la clase de interés.

Existen varias sutilezas con respecto al tamaño de las regiones. Por ejemplo, si se tiene un pico como el mostrado en medio de la figura 2.5, podría ser que haya pocos individuos asociados con los valores de las variables correspondientes al pico, en este caso, sólo los individuos con *variable* 1 = 3 y *variable* 2 = 3. Si el pico fuera más amplio, por ejemplo abarcando los rangos $1 \leq$ *variable* $1 \leq 4$ y $1 \leq$ *variable* $2 \leq 4$, entonces el número de individuos asociados con dicha región probablemente sería mayor. En ocasiones no importa solamente la altura de la región, sino también su forma y tamaño.

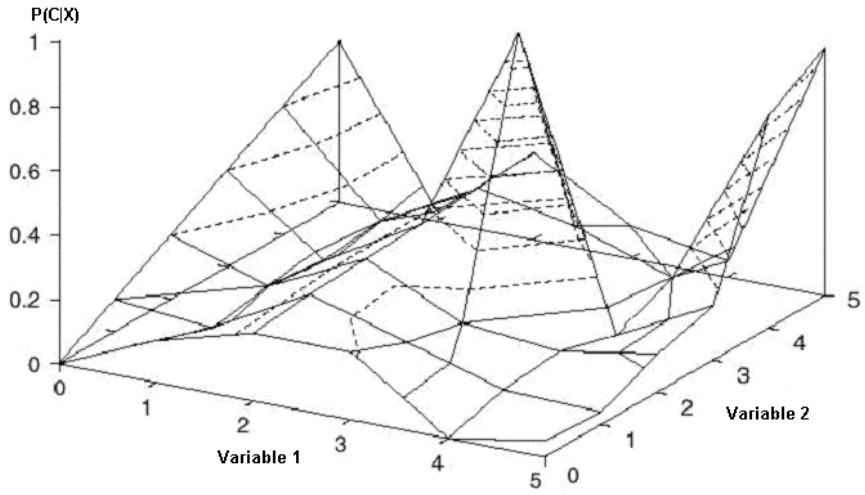

Fig. 2.5 Ejemplo de un paisaje de predictibilidad de dos variables métricas. La medida de predictibilidad usada es *p*(*C*|*X*) y la relación entre las variables no es lineal.

En problemas reales de minería de datos, generalmente sólo se cuenta con un número pequeño de muestras relativo al número total de posibilidades. Al conjunto que nos sirve para determinar la relación *F*(***X***) entre los valores de las variables de entrada, ***X*** y los valores conocidos de la variable de salida *F*, se le conoce como conjunto de entrenamiento. Con base en este conjunto, se deben inferir los valores de *F* para valores de ***X*** que no se conocían previamente. Esto se realiza mediante algún tipo de interpolación, la cual, de preferencia, debe ser más fiel en las regiones de mayor interés. Un modelo de minería de datos es una plantilla que se ajusta al paisaje de predictibilidad. Esta plantilla ofrece una manera de interpolar los valores conocidos de *F* de un individuo ***X***, para inferir los valores desconocidos de *F* de nuevos individuos ***X***.

A manera de ejemplo, considérese la figura 2.6, la cual muestra la relación entre los costos de un paciente en distintos años (1997 y 1998). Si se desea inferir los costos de un paciente para 1998 a partir de sus costos de 1997 existen varias opciones de modelos, pero en este sencillo ejemplo se puede elegir un modelo lineal.

Existe un gran número de modelos para resolver problemas de minería de datos, estos pueden ir desde modelos simples, como una regresión lineal o logística a modelos más

complejos, como las redes neuronales. La desventaja de los modelos simples es que, a no ser que capturen características inherentes en los datos, pueden tener malos desempeños. Sin embargo, su ventaja es que existen pocas probabilidades de que sobre-ajusten los datos. Por otra parte, los modelos más complejos ofrecen mayor flexibilidad, pero corren el riesgo de sobre-ajustar los datos. Entre más dispersión tengan los valores de la función de predictibilidad para un vector de características dado, se corre más riesgo de sobre-ajustar los datos. Si no se tiene mucha experiencia en reconocer y evitar el sobre-ajuste, lo mejor es evitar modelos sofisticados. El sobre-ajuste puede ser originado por una alta dispersión en la función de predictibilidad y por disponer de pocos datos. En términos generales, mientras más datos se tengan, entonces se pueden usar modelos que involucren más parámetros.

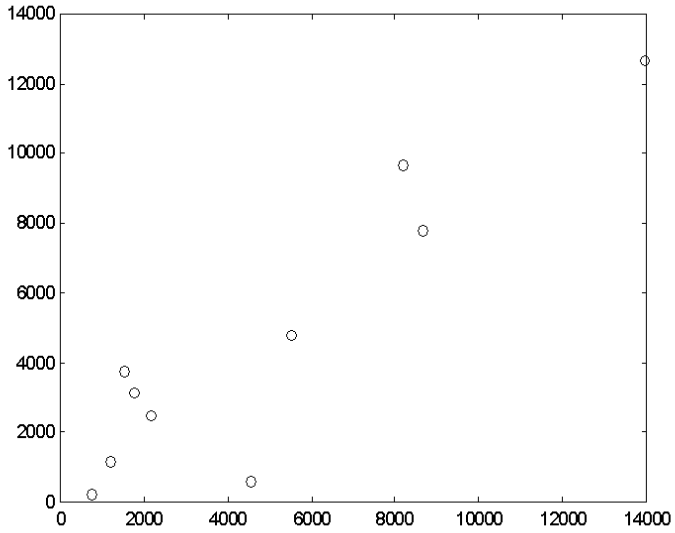

Fig. 2.6 Relación entre los costos de 1997 y 1998 para un conjunto de pacientes. Se busca predecir los costos de 1998 a partir de los costos de 1997.

Las mismas consideraciones se pueden aplicar a modelos donde se estima la distribución de probabilidad de la medida de predictibilidad directamente de los datos. Por ejemplo, cuando se construye $p(C|\boldsymbol{X})$ a partir de los datos. Se ha resaltado que generalmente en estos casos, el número de posibles vectores $\boldsymbol{X}$ es mucho mayor que el número de observaciones con que se cuenta. Para la gran mayoría de los vectores de características no se cuenta con observaciones, y en el caso de los vectores de características de los cuales sí hay observaciones generalmente son muy pocas, si no es que sólo una. En la sección 2.3 se presentó el *coarse graining* como un método para aminorar este problema, Mientras más grueso sea el *coarse graining* se obtiene más confiabilidad estadística. Sin embargo, esto dificulta la predicción a niveles más finos. Por ejemplo, dados $p(C|x_1{*})$ y $p(C|{*}\ x_2)$, ¿cómo se puede determinar $p(C|x_1\ x_2)$? Para determinar $p(C|x_1\ x_2)$ se deben hacer supuestos o aproximaciones. Un método robusto y poderoso para determinar las probabilidades $p(C|\boldsymbol{X})$ es la clasificación Bayesiana ingenua. El supuesto que se hace en este método es que las variables que componen al vector $\boldsymbol{X}$ son independientes. Este método ha mostrado ser muy útil tanto en problemas académicos como en aplicaciones del mundo real, aún cuando no se cumpla el supuesto de independencia de las variables.

2.5.2 Clasificación Bayesiana ingenua

La clasificación Bayesiana ingenua es un modelo fácil de implementar y es el método óptimo de aprendizaje supervisado cuando las variables de predicción son independientes dada una clase de membresía. Si las variables están correlacionadas entonces no es el método óptimo. Sin embargo, aún en estos casos, en la práctica, la clasificación Bayesiana ingenua ha resultado ser muy útil y difícil de mejorar sistemáticamente [6].

En el capítulo 1 se presentó el teorema de Bayes:

$$p(C \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid C)p(C)}{p(\boldsymbol{X})} \qquad \textbf{(2.5)}$$

En la clasificación Bayesiana ingenua se asume que las variables que componen al vector $\boldsymbol{X}$ son independientes, por lo que la probabilidad $p(\boldsymbol{X}|C)$ de la ecuación (2.5) se puede calcular como:

$$p(\boldsymbol{X} \mid C) = \prod_{i=1}^{N} p\left(x_{ij} \mid C\right) \qquad \textbf{(2.6)}$$

donde:

N es el número de variables que componen al vector $\boldsymbol{X}$.

$p(x_{ij}|C)$ es la probabilidad de tener el valor j-ésimo en la i-ésima variable, dada la clase C. Se calcula como el número de muestras que cumplen con la condición $x_i = valor_j$ y que además pertenecen a la clase C, entre el número de muestras que pertenecen a la clase C.

En el caso de que solamente se tengan dos clases, se puede asignar una puntuación a cada individuo mediante una función de aptitud, $f(\boldsymbol{X})$, definida mediante la siguiente ecuación:

$$f\left(\boldsymbol{X}\right) = log\,\frac{p(C_1 \mid \boldsymbol{X})}{p(C_2 \mid \boldsymbol{X})} \qquad \textbf{(2.7)}$$

Sustituyendo (2.5) en (2.7) se obtiene:

$$f\left(\boldsymbol{X}\right) = log\,\frac{p(\boldsymbol{X} \mid C_1)p(C_1)}{p(\boldsymbol{X} \mid C_2)p(C_2)} \qquad \textbf{(2.8)}$$

Manipulando (2.6) se obtiene:

$$f\left(\boldsymbol{X}\right) = log\,\frac{p(\boldsymbol{X} \mid C_1)p(C_1)}{p(\boldsymbol{X} \mid C_2)p(C_2)} = log\,\frac{p(\boldsymbol{X} \mid C_1)}{p(\boldsymbol{X} \mid C_2)} + log\,\frac{p(C_1)}{p(C_2)} \qquad \textbf{(2.9)}$$

Dado que $f(\boldsymbol{X})$ es una función de aptitud para la observación $\boldsymbol{X}$, se puede descartar el término constante para todo vector $\boldsymbol{X}$, $log\ p(C_1) / p(C_2)$ y obtener la siguiente función discriminante:

$$g(\boldsymbol{X}) = log\,\frac{p(\boldsymbol{X} \mid C_1)}{p(\boldsymbol{X} \mid C_2)} \qquad \textbf{(2.10)}$$

Sustituyendo (2.6) en (2.10) y manipulando se obtiene:

$$g(\boldsymbol{X}) = \sum_{i=1}^{N} log\,\frac{p(x_{ij} \mid C_1)}{p(x_{ij} \mid C_2)} \qquad \textbf{(2.11)}$$

De esta manera, en lugar de solamente asignar la clase C_1 o C_2 a un vector $\boldsymbol{X}$, se puede obtener una función de aptitud que entre mayor sea significa que el vector $\boldsymbol{X}$ tiene mayores probabilidades de pertenecer a la clase C_1 que a la clase C_2. Esta función de aptitud es muy útil porque no sólo sirve para asignar la clase a un individuo, representado por el vector $\boldsymbol{X}$, sino que también permite ordenar todos los individuos que se desean clasificar por la probabilidad de pertenencia a la clase C_1.

La clasificación Bayesiana ingenua trabaja muy bien, especialmente en conjuntos de datos de tamaño chico a medio, aún cuando las variables no sean independientes. Esto significa que generaliza bien sobre nuevos datos y que no sufre del problema de sobre-ajuste. La razón de esto podría ser que se tiene mayor confiabilidad estadística en el cálculo de $p(\boldsymbol{X}|C)$ porque se obtiene a partir de probabilidades $p(x_{ij}|C)$, y al calcular estas últimas probabilidades existen más observaciones para valores x_{ij}, que para un vector específico $\boldsymbol{X}$. Para dar un ejemplo, si cada variable puede tener 10 valores, entonces la probabilidad de tener una observación para un vector específico $\boldsymbol{X} = [\, x_{1j}, ..., x_{Nj} \,]$ es muy chica ($1/10^N$). Sin embargo, la probabilidad de tener una observación para un valor x_{ij} es mucho mayor (1/10), por lo que se tiene mayor confiabilidad estadística. Si las variables no son independientes y se desea considerar la dependencia de las variables en el cálculo de $p(\boldsymbol{X}|C)$, el desempeño que se puede ganar al tener una aproximación más exacta de $p(\boldsymbol{X}|C)$, se puede perder debido a la incertidumbre estadística originada por la falta de observaciones.

2.6 Enfoque multiperspectiva

No existe una receta mágica para resolver un problema de minería de datos. Como ya se ha mencionado, los problemas del mundo real tienen múltiples objetivos que requieren múltiples soluciones. Además, cada modelo de minería de datos tiene sus ventajas y desventajas. Debido a esto, no es conveniente asociar un único modelo con una sola medida de predictibilidad. Por ejemplo, los modelos de clasificación Bayesiana ingenua pueden ser muy buenos para predecir la pertenencia a una clase en datos con una distribución con poca densidad, pero no pueden descubrir las dependencias no lineales, potencialmente importantes, entre las diferentes variables. Es por esto que la minería de datos requiere un enfoque multiperspectiva. Lo más recomendable es que la multiperspectiva se aplique en varios niveles como son:

- *Tarea a realizar*. Generalmente para resolver problemas de minería se deben involucrar varias tareas como son distintos tipos de perfilación y de predicción. Por ejemplo, en la perfilación se puede usar ε y ε' para medir la importancia de indicadores particulares, o combinaciones de indicadores con respecto a alguna función objetivo, además ε' también se puede usar para seleccionar variables que mejoren los resultados de la predicción. También se pueden realizar distintas

predicciones, por ejemplo, además de predecir quienes serán los pacientes más costosos, quizá también se requiera predecir los costos asociados a esos pacientes.

- *Medida de predictibilidad*. Se pueden usar distintas medidas de predictibilidad como $p(C|\boldsymbol{X})$, $p(\boldsymbol{X}|C)$, $\varepsilon(\boldsymbol{X})$, o alguna otra función más específica al problema, como Costos($\boldsymbol{X}$), ROI($\boldsymbol{X}$), etc. El ROI (*Return of Investment*) es una medida del beneficio económico de un proyecto o inversión. Las medidas de predictibilidad se pueden combinar para robustecerlas o aminorar desventajas que pudieran tener.
- *Modelos a implementar*. Dependiendo de los objetivos (tiempo, transparencia, recursos, grado de acción, etc.) se pueden implementar distintos modelos. También influyen las características de los datos disponibles. Por ejemplo, quizás una regresión lineal de buenos resultados para un aspecto particular de un problema y después se encuentre que un algoritmo genético también proporciona buenos resultados y que combinando ambos modelos se mejoran los resultados obtenidos de manera independiente para cada modelo. O se puede usar una clasificación Bayesiana ingenua para cierta tarea de predicción y una red neuronal para otra tarea, tomando como entrada solamente las variables seleccionadas mediante ε'.
- *Tipos de variables*. Ejemplos: variables sociodemográficas versus historial de compra de un cliente; historial de condición médica versus historial de costos de pacientes. Dependiendo de la tarea a realizar se puede usar un solo tipo de datos (usar solo las variables de historial médico para predecir el estado de salud de un paciente) o se pueden usar los distintos tipos de datos disponibles (usar tanto las variables de condición médica como los costos actuales de un paciente para predecir sus costos en el siguiente año).
- *Escalas de tiempo*. Ejemplo: compras recientes versus historial de compras de varios años.
- *Escalas espaciales*. Ejemplo: variables sociodemográficas a nivel código postal versus variables sociodemográficas a nivel estado.
- *Acciones a realizar*. Se pueden realizar distintas acciones, que pueden ser desde acciones automatizadas (enviar un e-mail a un cliente potencial) hasta acciones que requieren la toma de una decisión de una persona (eliminar una campaña para atraer clientes que no está siendo efectiva).

Este enfoque multiperspectiva es la metodología que siguen las personas cuando afrontan una tarea compleja. En estos casos se requieren de distintas personas con distintas habilidades, en donde cada persona aporta sus puntos de vista y experiencia, enriqueciendo el trabajo. En lugar de dejar que las decisiones recaigan en una sola persona se toman decisiones de consenso, ya que éstas tienen más validez que las decisiones individuales.

Para facilitar el planteamiento del enfoque multiperspectiva considérese que cada componente uniperspectiva es un agente, donde la definición de agente es la siguiente:

"Un agente es un modelo que proporciona una opinión, perspectiva o predicción con base en datos de entrada".

Los ejemplos dados para los distintos niveles multiperspectiva corresponden a la definición de agentes: un agente puede ser un modelo de clasificación Bayesiana ingenua, un selector de variables mediante ε', un "agente de acción" que envía e-mails, etc. Una clase de agentes se puede dividir en subclases, por ejemplo, el agente de

clasificación Bayesiana ingenua se puede subdividir en agentes que usen distintos tipos de datos (historial médico, historial de costos, variables sociodemográficas, etc.).

El enfoque multiperspectiva origina las siguientes preguntas:

- ¿Cómo se mide el desempeño de cada agente? Dependiendo del agente se debe poder asociar una medida de desempeño. Esta medida puede ser de precisión, tiempo, robustez, etcétera. Mediante esta medida se pueden comparar los distintos agentes, siempre y cuando tenga sentido la comparación.
- ¿Cómo se generan los agentes? Se requiere analizar cuidadosamente el problema de minería de datos para determinar los mejores agentes a involucrar.
- ¿Cómo se integran la información de los distintos agentes? Existen muchas maneras de hacer esto, una manera común de integración es cuando la opinión de un agente se incluye para una tarea particular. Se debe buscar una manera de combinar las opiniones de los distintos agentes. En el contexto de clasificación, un individuo puede clasificarse mediante una regla de consenso. Esta regla asigna una calificación al individuo la cual mide la probabilidad de que el individuo pertenezca a la clase. Existen varias reglas de consenso para la clasificación de un individuo.

    - Ganador. Cada agente clasificador asigna una calificación al individuo de pertenencia a una clase. Gana el agente que le haya dado la calificación más alta al individuo. Por ejemplo, si un agente opinó que el individuo pertenece a la clase C_1 con una calificación mayor que otro agente, el cual opinó que pertenece a la clase C_2, entonces gana el primer agente y el individuo se asigna a la clase C_1. Hay que tener cuidado en normalizar las calificaciones para poder compararlas entre sí.
    - Promedio. Se promedian las calificaciones de los distintos agentes. Por ejemplo si dos agentes opinaron que el individuo pertenece a la clase C_2 y el promedio de calificación fue mayor que el promedio de tres agentes que opinaron que el individuo pertenece a la clase C_1, entonces el individuo se asigna a la clase C_2. Hay que tener cuidado en normalizar las calificaciones para poder compararlas entre sí.
    - Correspondencia. Es parecido a la regla promedio, pero en lugar de promediar sólo se cuenta el número de agentes. Por ejemplo si dos agentes opinaron que el individuo pertenece a la clase C_2 y otros tres agentes opinaron que el individuo pertenece a la clase C_1, entonces el individuo se asigna a la clase C_1.

En los siguientes dos capítulos se llevará a la práctica el enfoque multiperspectiva.

Capítulo 3: Caso de estudio "Predicción de compra de pólizas de seguro"

En el año 2000, el grupo CoIL (*The Computational Intelligence and Learning Cluster*) realizó el concurso *Challenge 2000* con la finalidad de:

- Promover la aplicación de inteligencia computacional y tecnología de aprendizaje a problemas del mundo real.
- Clarificar las relaciones entre diferentes enfoques.
- Estimular la búsqueda de soluciones que combinen diferentes métodos.

El objetivo del concurso CoIL *Challenge 2000* fue predecir y explicar la posesión de una póliza de seguro. La pregunta principal fue:

¿Puede predecir quienes serían las personas interesadas (20% del conjunto de prueba) en comprar una póliza de seguro para una casa rodante y dar una explicación de por qué?

La predicción se tenia que hacer sobre un conjunto de datos conocido como TIC (*The Insurance Company*) *Benchmark*, proporcionado por la compañía holandesa de minería de datos *Sentient Machine Research.* Este conjunto de datos está basado en aplicaciones de negocios reales y es apropiado para probar algoritmos de minería de datos o para usarse en sesiones de laboratorio de minería de datos.

El concurso fue dividido en dos categorías:

1. Predicción
2. Descripción

Para la tarea de predicción se les pidió a los participantes que identificaran, de los datos de prueba, el 20% (800 clientes) que contuvieran el mayor número de propietarios de póliza de casa rodante.

La tarea de descripción consistió en explicar por qué estos 800 clientes seleccionados eran los individuos más probables de comprar una póliza de casa rodante, es decir, se pedía proporcionar un entendimiento de las características de estos clientes.

El concurso registró a 147 participantes. De los cuales, 43 enviaron sus soluciones y 29 de los 43 participantes escribieron un reporte explicando sus métodos y resultados. Los participantes fueron investigadores, estudiantes, así como profesionales del área de minería de datos.

3.1 Descripción de los datos

Consisten en un conjunto de datos para el entrenamiento de los modelos de predicción y un conjunto de datos sobre los cuales se desea predecir.

3.1.1 Conjunto de datos para entrenar los modelos de predicción

Conjunto para entrenar y validar los modelos de predicción. Consiste de 5822 registros. Un registro por cliente. Cada registro consiste de 43 variables de datos sociodemográficos, 42 variables relacionadas con distintos tipos de pólizas que posee el cliente y la variable objetivo CARAVAN, que indica si un cliente posee una póliza de casa rodante. Los datos sociodemográficos se obtuvieron a partir de los códigos postales, así que todos los clientes que viven en áreas con el mismo código postal tienen las mismos variables sociodemográficas. En total son 86 variables, incluyendo la variable a predecir, CARAVAN.

Las variables relacionadas con distintos tipos de pólizas que posee el cliente son básicamente de dos tipos:

1. Aquéllas que se refieren al nivel de contribución del cliente para cada tipo de póliza (21 variables).
2. Aquéllas que se refieren al número de pólizas que tiene el cliente de cada tipo de póliza (21 variables).

3.1.2 Conjunto de datos en los cuales se desea predecir

Conjunto de datos para predecir. Consiste de 4000 registros de clientes. Tiene el mismo formato que el conjunto anterior, pero no tiene la variable objetivo, por lo que son 85 variables en lugar de 86. La variable objetivo se proporcionó por separado después de que los participantes enviaron sus resultados y solamente se usó para medir el desempeño de las predicciones.

Ambos conjuntos están en formato delimitado por tabuladores.

Los datos de entrenamiento y de prueba fueron tomados aleatoriamente de la misma población. Para cada cliente se tenían los valores de las 85 variables de los datos (como ya se mencionó, el conjunto de entrenamiento tenía adicionalmente la variable objetivo).

De los 4000 clientes del conjunto de prueba, los participantes tenían que identificar un conjunto de 800 clientes con mayor probabilidad de tener pólizas de casa rodante.

3.1.3 Variables de los datos

Las variables que se tienen de los datos son las siguientes:

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
1	MOSTYPE	Subtipo de cliente	No	41	L0
2	MAANTHUI	Número de casas	Sí	10	Entre 1 y 10
3	MGEMOMV	Habitantes promedio por casa	Sí	6	Entre 1 y 6
4	MGEMLEEF	Edad promedio	Sí	6	L1
5	MOSHOOFD	Tipo principal de cliente	No	10	L2
6	MGODRK	Católico romano	Sí	10	L3
7	MGODPR	Protestante	Sí	10	L3
8	MGODOV	Otra religión	Sí	10	L3

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
9	MGODGE	Sin religión	Sí	10	L3
10	MRELGE	Casado	Sí	10	L3
11	MRELSA	Unión libre	Sí	10	L3
12	MRELOV	Otra relación	Sí	10	L3
13	MFALLEEN	Soltero	Sí	10	L3
14	MFGEKIND	Familia con hijos	Sí	10	L3
15	MFWEKIND	Familia sin hijos	Sí	10	L3
16	MOPLHOOG	Educación de nivel alto	Sí	10	L3
17	MOPLMIDD	Educación de nivel medio	Sí	10	L3
18	MOPLLAAG	Educación de nivel bajo	Sí	10	L3
19	MBERHOOG	Estatus alto	Sí	10	L3
20	MBERZELF	Empresario	Sí	10	L3
21	MBERBOER	Granjero	Sí	10	L3
22	MBERMIDD	Administración media	Sí	10	L3
23	MBERARBG	Trabajadores capacitados	Sí	10	L3
24	MBERARBO	Trabajadores no capacitados	Sí	10	L3
25	MSKA	Clase social A	Sí	10	L3
26	MSKB1	Clase social B1	Sí	10	L3
27	MSKB2	Clase social B2	Sí	10	L3
28	MSKC	Clase social C	Sí	10	L3
29	MSKD	Clase social D	Sí	10	L3
30	MHHUUR	Casa rentada	Sí	10	L3
31	MHKOOP	Casa propia	Sí	10	L3
32	MAUT1	1 auto	Sí	10	L3
33	MAUT2	2 autos	Sí	10	L3
34	MAUT0	Sin auto	Sí	10	L3
35	MZFONDS	Servicio nacional de salud	Sí	10	L3
36	MZPART	Segura privado de salud	Sí	10	L3
37	MINKM30	Ingresos < 30,000	Sí	10	L3
38	MINK3045	Ingresos 30 - 45,000	Sí	10	L3
39	MINK4575	Ingresos 45 - 75,000	Sí	10	L3
40	MINK7512	Ingresos 75 - 122,000	Sí	10	L3
41	MINK123M	Ingresos > 123,000	Sí	10	L3
42	MINKGEM	Ingreso promedio	Sí	10	L3
43	MKOOPKLA	Clase con poder adquisitivo	Sí	10	L3
44	PWAPART	Contribución de seguros contra daño a terceros	Sí	10	L4
45	PWABEDR	Contribución de seguros contra daño a terceros (empresas)	Sí	10	L4
46	PWALAND	Contribución de seguros contra daño a terceros (agricultura)	Sí	10	L4
47	PPERSAUT	Contribución de pólizas de auto	Sí	10	L4
48	PBESAUT	Contribución de pólizas de camioneta repartidora	Sí	10	L4
49	PMOTSCO	Contribución de pólizas de motocicleta	Sí	10	L4

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
50	PVRAAUT	Contribución de pólizas de camión de carga	Sí	10	L4
51	PAANHANG	Contribución de pólizas de tráiler	Sí	10	L4
52	PTRACTOR	Contribución de pólizas de tractor	Sí	10	L4
53	PWERKT	Contribución de pólizas de máquinas para agricultura	Sí	10	L4
54	PBROM	Contribución de pólizas de ciclomotor	Sí	10	L4
55	PLEVEN	Contribución de seguros de vida	Sí	10	L4
56	PPERSONG	Contribución de pólizas de seguro contra accidentes	Sí	10	L4
57	PGEZONG	Contribución de pólizas de seguro contra accidentes familiares	Sí	10	L4
58	PWAOREG	Contribución de pólizas de seguro contra discapacidades	Sí	10	L4
59	PBRAND	Contribución de pólizas contra incendio	Sí	10	L4
60	PZEILPL	Contribución de pólizas de tablas de surfeo	Sí	10	L4
61	PPLEZIER	Contribución de pólizas de bote	Sí	10	L4
62	PFIETS	Contribución de pólizas de bicicleta	Sí	10	L4
63	PINBOED	Contribución de pólizas de seguro de propiedad	Sí	10	L4
64	PBYSTAND	Contribución de pólizas de seguridad social	Sí	10	L4
65	AWAPART	Número de seguros contra daño a terceros	Sí	13	Entre 0 y 12
66	AWABEDR	Número de seguros contra daño a terceros (empresas)	Sí	13	Entre 0 y 12
67	AWALAND	Número de seguros contra daño a terceros (agricultura)	Sí	13	Entre 0 y 12
68	APERSAUT	Número de pólizas de auto	Sí	13	Entre 0 y 12
69	ABESAUT	Número de pólizas de camioneta repartidora	Sí	13	Entre 0 y 12
70	AMOTSCO	Número de pólizas de motocicleta	Sí	13	Entre 0 y 12
71	AVRAAUT	Número de pólizas de camión de carga	Sí	13	Entre 0 y 12
72	AAANHANG	Número de pólizas de tráiler	Sí	13	Entre 0 y 12
73	ATRACTOR	Número de pólizas de tractor	Sí	13	Entre 0 y 12
74	AWERKT	Número de pólizas de máquinas para agricultura	Sí	13	Entre 0 y 12
75	ABROM	Número de pólizas de ciclomotor	Sí	13	Entre 0 y 12
76	ALEVEN	Número de seguros de vida	Sí	13	Entre 0 y 12

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
77	APERSONG	Número de pólizas de seguro contra accidentes	Sí	13	Entre 0 y 12
78	AGEZONG	Número de pólizas de seguro contra accidentes familiares	Sí	13	Entre 0 y 12
79	AWAOREG	Número de pólizas de seguro contra discapacidades	Sí	13	Entre 0 y 12
80	ABRAND	Número de pólizas contra incendio	Sí	13	Entre 0 y 12
81	AZEILPL	Número de pólizas de tablas de surfeo	Sí	13	Entre 0 y 12
82	APLEZIER	Número de pólizas de bote	Sí	13	Entre 0 y 12
83	AFIETS	Número de pólizas de bicicleta	Sí	13	Entre 0 y 12
84	AINBOED	Número de pólizas de seguro de propiedad	Sí	13	Entre 0 y 12
85	ABYSTAND	Número de pólizas de seguridad social	Sí	13	Entre 0 y 12
86	CARAVAN	Número de pólizas de casa rodante	Sí	2	0 ó 1

Tabla 3.1 Variables de los datos CoIL.

La columna *Métrica* indica si la variable tiene valores que tienen un orden entre sí. Se observa que sólo hay dos variables que no son métricas ya que representan valores categóricos.

La columna *Cardinalidad* indica cuántos posibles valores puede tener la variable.

La variable CARAVAN es la variable objetivo. Para el conjunto de prueba esta variable se proporcionó por separado con la finalidad de que sólo se usara para evaluar las predicciones realizadas.

Los posibles valores de las variables se muestran en el Apéndice A.

3.2 Análisis inicial de los datos

El resumen de los datos de entrenamiento y prueba se muestra en la siguiente tabla:

Variable	Datos de entrenamiento				Datos de prueba			
	Valor mínimo	Valor máximo	Media	Desviación Estándar	Valor mínimo	Valor máximo	Media	Desviación Estándar
MOSTYPE	1	41	24.25	12.85	1	41	24.25	13.02
MAANTHUI	1	10	1.11	0.41	1	10	1.11	0.42
MGEMOMV	1	5	2.68	0.79	1	6	2.68	0.77
MGEMLEEF	1	6	2.99	0.81	1	6	3.00	0.79
MOSHOOFD	1	10	5.77	2.86	1	10	5.79	2.90
MGODRK	0	9	0.70	1.00	0	9	0.71	1.03
MGODPR	0	9	4.63	1.72	0	9	4.65	1.73
MGODOV	0	5	1.07	1.02	0	5	1.02	1.00
MGODGE	0	9	3.26	1.60	0	9	3.27	1.62

Variable	Datos de entrenamiento				Datos de prueba			
	Valor mínimo	Valor máximo	Media	Desviación Estándar	Valor mínimo	Valor máximo	Media	Desviación Estándar
MRELGE	0	9	6.18	1.91	0	9	6.20	1.88
MRELSA	0	7	0.88	0.97	0	7	0.86	0.96
MRELOV	0	9	2.29	1.72	0	9	2.28	1.69
MFALLEEN	0	9	1.89	1.80	0	9	1.89	1.75
MFGEKIND	0	9	3.23	1.62	0	9	3.25	1.59
MFWEKIND	0	9	4.30	2.01	0	9	4.31	1.95
MOPLHOOG	0	9	1.46	1.62	0	9	1.52	1.68
MOPLMIDD	0	9	3.35	1.76	0	9	3.24	1.67
MOPLLAAG	0	9	4.57	2.30	0	9	4.62	2.25
MBERHOOG	0	9	1.90	1.80	0	9	1.90	1.84
MBERZELF	0	5	0.40	0.78	0	5	0.41	0.80
MBERBOER	0	9	0.52	1.06	0	9	0.58	1.17
MBERMIDD	0	9	2.90	1.84	0	9	2.85	1.86
MBERARBG	0	9	2.22	1.73	0	9	2.24	1.77
MBERARBO	0	9	2.31	1.69	0	9	2.27	1.67
MSKA	0	9	1.62	1.72	0	9	1.69	1.77
MSKB1	0	9	1.61	1.33	0	9	1.58	1.31
MSKB2	0	9	2.20	1.53	0	9	2.21	1.54
MSKC	0	9	3.76	1.94	0	9	3.72	1.96
MSKD	0	9	1.07	1.30	0	8	1.07	1.29
MHHUUR	0	9	4.24	3.09	0	9	4.12	3.10
MHKOOP	0	9	4.77	3.09	0	9	4.89	3.10
MAUT1	0	9	6.04	1.55	0	9	6.00	1.53
MAUT2	0	7	1.32	1.20	0	9	1.36	1.23
MAUT0	0	9	1.96	1.60	0	9	1.95	1.59
MZFONDS	0	9	6.28	1.98	0	9	6.22	2.03
MZPART	0	9	2.73	1.98	0	9	2.78	2.03
MINKM30	0	9	2.57	2.09	0	9	2.58	2.05
MINK3045	0	9	3.54	1.88	0	9	3.46	1.85
MINK4575	0	9	2.73	1.93	0	9	2.75	1.98
MINK7512	0	9	0.80	1.16	0	9	0.83	1.19
MINK123M	0	9	0.20	0.55	0	6	0.22	0.58
MINKGEM	0	9	3.78	1.32	0	9	3.83	1.35
MKOOPKLA	1	8	4.24	2.01	1	8	4.30	1.99
PWAPART	0	3	0.77	0.96	0	3	0.76	0.95
PWABEDR	0	6	0.04	0.36	0	6	0.04	0.35
PWALAND	0	4	0.07	0.50	0	4	0.08	0.52
PPERSAUT	0	8	2.97	2.92	0	9	2.94	2.92
PBESAUT	0	7	0.05	0.53	0	7	0.06	0.61
PMOTSCO	0	7	0.18	0.90	0	6	0.16	0.88
PVRAAUT	0	9	0.01	0.24	0	7	0.01	0.23
PAANHANG	0	5	0.02	0.21	0	3	0.02	0.18
PTRACTOR	0	6	0.09	0.60	0	7	0.10	0.61
PWERKT	0	6	0.01	0.23	0	6	0.01	0.19
PBROM	0	6	0.22	0.81	0	6	0.22	0.81

Variable	Datos de entrenamiento				Datos de prueba			
	Valor mínimo	Valor máximo	Media	Desviación Estándar	Valor mínimo	Valor máximo	Media	Desviación Estándar
PLEVEN	0	9	0.19	0.90	0	7	0.21	0.93
PPERSONG	0	6	0.01	0.21	0	5	0.01	0.15
PGEZONG	0	3	0.02	0.19	0	3	0.02	0.24
PWAOREG	0	7	0.02	0.38	0	7	0.02	0.38
PBRAND	0	8	1.83	1.88	0	8	1.88	1.88
PZEILPL	0	3	0.00	0.04	0	2	0.00	0.07
PPLEZIER	0	6	0.02	0.27	0	5	0.01	0.19
PFIETS	0	1	0.03	0.16	0	1	0.03	0.16
PINBOED	0	6	0.02	0.20	0	6	0.02	0.22
PBYSTAND	0	5	0.05	0.41	0	5	0.04	0.38
AWAPART	0	2	0.40	0.49	0	2	0.40	0.49
AWABEDR	0	5	0.01	0.13	0	1	0.01	0.11
AWALAND	0	1	0.02	0.14	0	1	0.02	0.15
APERSAUT	0	7	0.56	0.60	0	12	0.55	0.61
ABESAUT	0	4	0.01	0.13	0	5	0.01	0.13
AMOTSCO	0	8	0.04	0.23	0	3	0.04	0.22
AVRAAUT	0	3	0.00	0.06	0	4	0.00	0.08
AAANHANG	0	3	0.01	0.13	0	2	0.01	0.10
ATRACTOR	0	4	0.03	0.24	0	6	0.04	0.26
AWERKT	0	6	0.01	0.12	0	4	0.00	0.09
ABROM	0	2	0.07	0.27	0	3	0.07	0.27
ALEVEN	0	8	0.08	0.38	0	5	0.08	0.39
APERSONG	0	1	0.01	0.07	0	1	0.00	0.06
AGEZONG	0	1	0.01	0.08	0	1	0.01	0.10
AWAOREG	0	2	0.00	0.08	0	1	0.00	0.06
ABRAND	0	7	0.57	0.56	0	6	0.58	0.56
AZEILPL	0	1	0.00	0.02	0	1	0.00	0.04
APLEZIER	0	2	0.01	0.08	0	2	0.00	0.07
AFIETS	0	3	0.03	0.21	0	4	0.03	0.21
AINBOED	0	2	0.01	0.09	0	1	0.01	0.10
ABYSTAND	0	2	0.01	0.12	0	1	0.01	0.11
CARAVAN	0	1	0.06	0.24				

Tabla 3.2 Principales estadisticas de los datos de entrenamiento y prueba.

En los datos se observa que las principales estadísticas entre los dos conjuntos de datos son similares, lo cual era de esperarse ya que ambas muestras se extrajeron del mismo conjunto de datos. En la figura 3.1 se aprecian más claramente las similitudes entre las variables de ambos conjuntos. En dicha figura se muestra una gráfica de dispersión de las medias de las variables de ambos conjuntos de datos, entrenamiento y prueba. En el eje X se grafican las medias de todas las variables de los datos de entrenamiento (con excepción de la variable CARAVAN). En el eje Y se grafican las medias de todas las variables de los datos de prueba.

La figura 3.2 es un acercamiento de la figura 3.1. Se removió el punto más lejano para poder ver con más detalle los puntos entre cero y siete. Al igual que con la gráfica 3.1, se observa que los datos son similares con respecto a sus medias.

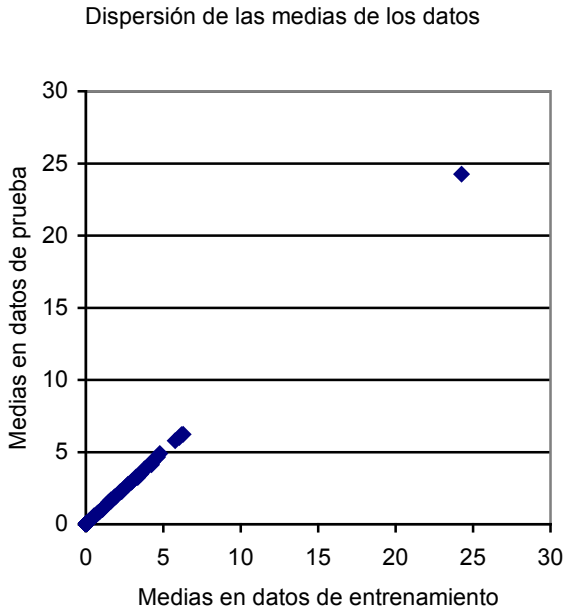

Fig. 3.1 Comparación de variables de entrenamiento versus variables de prueba.

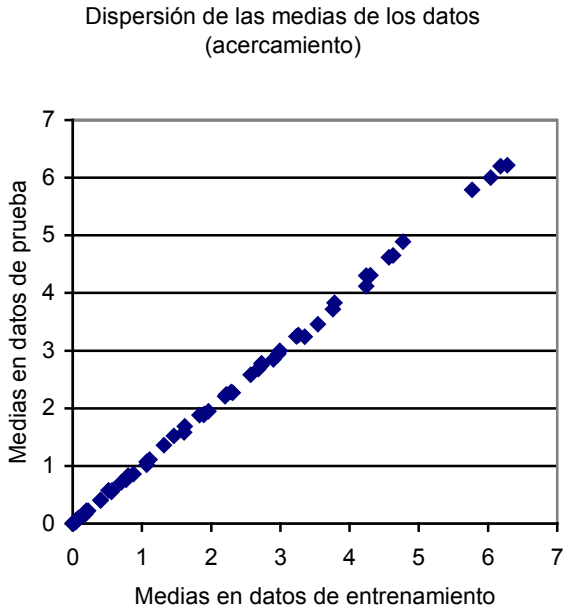

Fig. 3.2 Comparación de variables de entrenamiento versus variables de prueba (acercamiento).

Otro análisis inicial de los datos es obtener las tablas de frecuencias para cada variable. Es decir, obtener una tabla que indique cuantas individuos hay para posible valor de todas las variables. Esto nos da una idea general del comportamiento de las variables.

Para analizar las variables, éstas se pueden dividir en tres grupos:

1. 43 Variables sociodemográficas.
2. 21 Variables del nivel de contribución del cliente (una variable por tipo de póliza).
3. 21 Variables del número de pólizas del cliente (una variable por tipo de póliza).

Variables sociodemográficas (todos los datos)

Las variables sociodemográficas proporcionan información sobre la zona de residencia de los individuos: a un mismo código postal le corresponden los mismos valores de variables sociodemográficas. La desventaja de esto es que la información no es a nivel de individuo, lo cual sería lo ideal ya que estamos interesados en conocer características de los individuos, no de las zonas donde viven, pero de todas maneras estas variables nos proporcionan información general sobre los individuos. Además, como se mencionó en el capítulo 2, el enfoque multiperspectiva se puede aplicar en distintos niveles. Un nivel es por tipo de variable. En este caso, las variables sociodemográficas se pueden emplear para seleccionar nuevos clientes, cosa que no se puede hacer con variables de historial de los clientes.

Las siguientes cuatro tablas muestran las frecuencias de los distintos valores de las variables sociodemográficas. Debido a que la variable MOSTYPE tiene 41 posibles valores, se muestran sus frecuencias en tablas separadas al resto de las 42 variables sociodemográficas, ya que estas últimas sólo tienen hasta 10 posibles valores.

Variable	Frecuencias de los distintos valores de MOSTYPE									
MOSTYPE	1	2	3	4	5	6	7	8	9	10
	124	82	249	52	45	119	44	339	278	165
	11	12	13	14	15	16	17	18	19	20
	153	111	179	0	5	16	9	19	3	25
	21	22	23	24	25	26	27	28	29	30
	15	98	251	180	82	48	50	25	86	118
	31	32	33	34	35	36	37	38	39	40
	205	141	810	182	214	225	132	339	328	71
	41									
	205									

Tabla 3.3 Frecuencias de los distintos valores de MOSTYPE.

Variable	Frecuencias de los distintos valores de 42 variables sociodemográficas										
	0	1	2	3	4	5	6	7	8	9	10
MAANTHUI		5267	505	39	2	1	1	5	1	0	1
MGEMOMV		284	2131	2646	693	68	0				
MGEMLEEF		74	1452	3000	1073	193	30				
MOSHOOFD		552	502	886	52	569	205	550	1563	667	276
MGODRK	3228	1599	733	152	66	18	13	6	3	4	

Variable	Frecuencias de los distintos valores de 42 variables sociodemográficas										
	0	1	2	3	4	5	6	7	8	9	10
MGODPR	78	134	396	590	1607	1501	714	564	65	173	
MGODOV	2003	2014	1388	257	132	28	0	0	0	0	
MGODGE	456	230	1055	1453	1334	963	217	101	5	8	
MRELGE	64	75	157	246	324	946	1172	1683	361	794	
MRELSA	2448	2030	1075	159	78	18	13	1	0	0	
MRELOV	1173	539	1756	1152	648	266	179	64	21	24	
MFALLEEN	1757	951	1247	848	519	259	127	67	24	23	
MFGEKIND	371	372	1060	1498	1455	606	321	96	14	29	
MFWEKIND	153	292	635	973	1137	1106	783	351	206	186	
MOPLHOOG	2147	1322	1144	547	326	187	67	51	22	9	
MOPLMIDD	423	383	937	1330	1426	738	348	157	37	43	
MOPLLAAG	299	243	667	680	851	1009	856	640	254	323	
MBERHOOG	1524	1245	1364	756	397	249	138	92	26	31	
MBERZELF	4171	1202	348	37	12	52	0	0	0	0	
MBERBOER	4176	854	487	143	77	59	14	3	5	4	
MBERMIDD	667	403	1491	1394	953	431	211	178	14	80	
MBERARBG	1167	921	1382	1167	604	310	169	68	24	10	
MBERARBO	968	980	1439	1109	772	331	122	66	9	26	
MSKA	1738	1569	1198	685	261	127	96	79	13	56	
MSKB1	1353	1480	1783	775	298	78	25	5	8	17	
MSKB2	990	861	1676	1175	652	357	96	6	7	2	
MSKC	364	272	870	1090	1159	1168	487	217	71	124	
MSKD	2607	1563	852	441	223	100	22	13	0	1	
MHHUUR	949	428	717	593	517	519	382	425	532	760	
MHKOOP	760	530	426	382	499	520	604	724	428	949	
MAUT1	19	14	58	231	448	1210	1663	1413	261	505	
MAUT2	1854	1468	1748	385	301	56	9	1	0	0	
MAUT0	1450	776	1625	1066	587	174	89	25	13	17	
MZFONDS	55	15	307	177	357	974	875	1511	699	852	
MZPART	852	699	1511	849	992	364	178	307	15	55	
MINKM30	1304	630	1094	1079	599	568	293	156	48	51	
MINK3045	465	268	919	1147	1356	931	406	205	35	90	
MINK4575	891	657	1165	1215	1034	498	125	93	53	91	
MINK7512	3246	1359	736	246	147	71	8	1	4	4	
MINK123M	4900	763	96	36	24	1	0	1	0	1	
MINKGEM	25	49	651	1932	1854	733	355	131	70	22	
MKOOPKLA	0	587	425	1524	902	583	901	474	426	0	

Tabla 3.4 Frecuencias de los distintos valores de 42 variables sociodemográficas.

Variable	Porcentajes de los distintos valores de MOSTYPE									
MOSTYPE	1	2	3	4	5	6	7	8	9	10
	2.13	1.41	4.28	0.89	0.77	2.04	0.76	5.82	4.77	2.83
	11	12	13	14	15	16	17	18	19	20
	2.63	1.91	3.07	0.00	0.09	0.27	0.15	0.33	0.05	0.43
	21	22	23	24	25	26	27	28	29	30
	0.26	1.68	4.31	3.09	1.41	0.82	0.86	0.43	1.48	2.03
	31	32	33	34	35	36	37	38	39	40
	3.52	2.42	13.91	3.13	3.68	3.86	2.27	5.82	5.63	1.22
	41									
	3.52									

Tabla 3.5 Frecuencias normalizadas de los distintos valores de MOSTYPE.

Variable	Porcentajes de los distintos valores de 42 variables sociodemográficas										
	0	1	2	3	4	5	6	7	8	9	10
MAANTHUI		90.47	8.67	0.67	0.03	0.02	0.02	0.09	0.02	0.00	0.02
MGEMOMV		4.88	36.60	45.45	11.90	1.17	0.00				
MGEMLEEF		1.27	24.94	51.53	18.43	3.32	0.52				
MOSHOOFD		9.48	8.62	15.22	0.89	9.77	3.52	9.45	26.85	11.46	4.74
MGODRK	55.44	27.46	12.59	2.61	1.13	0.31	0.22	0.10	0.05	0.07	
MGODPR	1.34	2.30	6.80	10.13	27.60	25.78	12.26	9.69	1.12	2.97	
MGODOV	34.40	34.59	23.84	4.41	2.27	0.48	0.00	0.00	0.00	0.00	
MGODGE	7.83	3.95	18.12	24.96	22.91	16.54	3.73	1.73	0.09	0.14	
MRELGE	1.10	1.29	2.70	4.23	5.57	16.25	20.13	28.91	6.20	13.64	
MRELSA	42.05	34.87	18.46	2.73	1.34	0.31	0.22	0.02	0.00	0.00	
MRELOV	20.15	9.26	30.16	19.79	11.13	4.57	3.07	1.10	0.36	0.41	
MFALLEEN	30.18	16.33	21.42	14.57	8.91	4.45	2.18	1.15	0.41	0.40	
MFGEKIND	6.37	6.39	18.21	25.73	24.99	10.41	5.51	1.65	0.24	0.50	
MFWEKIND	2.63	5.02	10.91	16.71	19.53	19.00	13.45	6.03	3.54	3.19	
MOPLHOOG	36.88	22.71	19.65	9.40	5.60	3.21	1.15	0.88	0.38	0.15	
MOPLMIDD	7.27	6.58	16.09	22.84	24.49	12.68	5.98	2.70	0.64	0.74	
MOPLLAAG	5.14	4.17	11.46	11.68	14.62	17.33	14.70	10.99	4.36	5.55	
MBERHOOG	26.18	21.38	23.43	12.99	6.82	4.28	2.37	1.58	0.45	0.53	
MBERZELF	71.64	20.65	5.98	0.64	0.21	0.89	0.00	0.00	0.00	0.00	
MBERBOER	71.73	14.67	8.36	2.46	1.32	1.01	0.24	0.05	0.09	0.07	
MBERMIDD	11.46	6.92	25.61	23.94	16.37	7.40	3.62	3.06	0.24	1.37	
MBERARBG	20.04	15.82	23.74	20.04	10.37	5.32	2.90	1.17	0.41	0.17	
MBERARBO	16.63	16.83	24.72	19.05	13.26	5.69	2.10	1.13	0.15	0.45	
MSKA	29.85	26.95	20.58	11.77	4.48	2.18	1.65	1.36	0.22	0.96	
MSKB1	23.24	25.42	30.63	13.31	5.12	1.34	0.43	0.09	0.14	0.29	
MSKB2	17.00	14.79	28.79	20.18	11.20	6.13	1.65	0.10	0.12	0.03	
MSKC	6.25	4.67	14.94	18.72	19.91	20.06	8.36	3.73	1.22	2.13	
MSKD	44.78	26.85	14.63	7.57	3.83	1.72	0.38	0.22	0.00	0.02	
MHHUUR	16.30	7.35	12.32	10.19	8.88	8.91	6.56	7.30	9.14	13.05	

Variable	Porcentajes de los distintos valores de 42 variables sociodemográficas										
	0	1	2	3	4	5	6	7	8	9	10
MAANTHUI		90.47	8.67	0.67	0.03	0.02	0.02	0.09	0.02	0.00	0.02
MHKOOP	13.05	9.10	7.32	6.56	8.57	8.93	10.37	12.44	7.35	16.30	
MAUT1	0.33	0.24	1.00	3.97	7.69	20.78	28.56	24.27	4.48	8.67	
MAUT2	31.84	25.21	30.02	6.61	5.17	0.96	0.15	0.02	0.00	0.00	
MAUT0	24.91	13.33	27.91	18.31	10.08	2.99	1.53	0.43	0.22	0.29	
MZFONDS	0.94	0.26	5.27	3.04	6.13	16.73	15.03	25.95	12.01	14.63	
MZPART	14.63	12.01	25.95	14.58	17.04	6.25	3.06	5.27	0.26	0.94	
MINKM30	22.40	10.82	18.79	18.53	10.29	9.76	5.03	2.68	0.82	0.88	
MINK3045	7.99	4.60	15.78	19.70	23.29	15.99	6.97	3.52	0.60	1.55	
MINK4575	15.30	11.28	20.01	20.87	17.76	8.55	2.15	1.60	0.91	1.56	
MINK7512	55.75	23.34	12.64	4.23	2.52	1.22	0.14	0.02	0.07	0.07	
MINK123M	84.16	13.11	1.65	0.62	0.41	0.02	0.00	0.02	0.00	0.02	
MINKGEM	0.43	0.84	11.18	33.18	31.84	12.59	6.10	2.25	1.20	0.38	
MKOOPKLA	0.00	10.08	7.30	26.18	15.49	10.01	15.48	8.14	7.32	0.00	

Tabla 3.6 Frecuencias normalizadas de los distintos valores de 42 variables sociodemográficas.

Analizando las tablas de frecuencias se observa lo siguiente:

Variable	Lo qué nos dice la variable
MOSTYPE	14% vive en zonas con familias grandes de clase baja (MOSTYPE = 33) 6% vive en zonas con familias de clase media (MOSTYPE = 8)
MAANTHUI	90% vive en zonas donde las personas tienen una casa (MAANTHUI = 1)
MGEMOMV	45% vive en zonas con 3 habitantes por casa (MGEMOMV = 3)
MGEMLEEF	52% vive en zonas donde la edad promedio está entre 40 y 50 años (MGEMLEEF = 3)
MOSHOOFD	27% vive en zonas que tienen familias con adultos (MOSHOOFD = 8)
MGODRK	99% vive en zonas donde el porcentaje de católicos es menor al 50% 55% vive en zonas donde el porcentaje de católicos es 0%
MGODPR	48% vive en zonas donde el porcentaje de protestantes es menor al 50%
MGODOV	99% vive en zonas donde el porcentaje de personas con otra religión es menor al 50% 34% vive en zonas donde el porcentaje de personas que no tienen otra religión es 0%
MGODGE	78% vive en zonas donde el porcentaje de personas sin religión es menor al 50%
MRELGE	15% vive en zonas donde el porcentaje de casados es menor al 50%
MRELSA	99% vive en zonas donde el porcentaje de personas en unión libre es menor al 50% 42% vive en zonas donde el porcentaje de personas en unión libre es 0%
MRELOV	90% vive en zonas donde el porcentaje de personas con otro tipo de relación es menor al 50%
MFALLEEN	91% vive en zonas donde el porcentaje de solteros es menor al 50%
MFGEKIND	82% vive en zonas donde el porcentaje de familias sin hijos es menor al 50%
MFWEKIND	55% vive en zonas donde el porcentaje de familias con hijos es menor al 55%
MOPLHOOG	94% vive en zonas donde el porcentaje de nivel educativo alto es menor al 50%
MOPLMIDD	77% vive en zonas donde el porcentaje de nivel educativo medio es menor al 50%

Variable	Lo qué nos dice la variable
MOPLLAAG	47% vive en zonas donde el porcentaje de nivel educativo bajo es menor al 50%
MBERHOOG	91% vive en zonas donde el porcentaje de un estatus alto es menor al 50%
MBERZELF	99% vive en zonas donde el porcentaje de empresarios es menor al 50%
MBERBOER	99% vive en zonas donde el porcentaje de granjeros es menor al 50%
MBERMIDD	84% vive en zonas donde el porcentaje de administradores es menor al 50%
MBERARBG	90% vive en zonas donde el porcentaje de trabajadores capacitados es menor al 50%
MBERARBO	90% vive en zonas donde el porcentaje de trabajadores no capacitados es menor al 50%
MSKA	94% vive en zonas donde el porcentaje de personas de clase social A es menor al 50%
MSKB1	98% vive en zonas donde el porcentaje de personas de clase social B1 es menor al 50%
MSKB2	92% vive en zonas donde el porcentaje de personas de clase social B2 es menor al 50%
MSKC	65% vive en zonas donde el porcentaje de personas de clase social C es menor al 50%
MSKD	98% vive en zonas donde el porcentaje de personas de clase social D es menor al 50%
MHHUUR	55% vive en zonas donde el porcentaje de personas que rentan casa es menor al 50%
MHKOOP	45% vive en zonas donde el porcentaje de personas con casa propia es menor al 50%
MAUT1	13% vive en zonas donde el porcentaje de personas con 1 auto es menor al 50%
MAUT2	99% vive en zonas donde el porcentaje de personas con 2 autos es menor al 50%
MAUT0	95% vive en zonas donde el porcentaje de personas sin auto es menor al 50%
MZFONDS	16% vive en zonas donde el porcentaje de personas con servicio nacional de salud es menor al 50%
MZPART	84% vive en zonas donde el porcentaje de personas con servicio privado de salud es menor al 50%
MINKM30	81% vive en zonas donde el porcentaje de personas con ingresos menores a 30,000 florines es menor al 50%
MINK3045	71% vive en zonas donde el porcentaje de personas con ingresos entre 30,000 y 45,000 florines es menor al 50%
MINK4575	85% vive en zonas donde el porcentaje de personas con ingresos entre 45,000 y 75,000 florines es menor al 50%
MINK7512	98% vive en zonas donde el porcentaje de personas con ingresos entre 75,000 y 122,000 florines es menor al 50%
MINK123M	100% vive en zonas donde el porcentaje de personas con ingresos mayores a 123,000 florines es menor al 50%
MINKGEM	77% vive en zonas donde el porcentaje de personas con ingresos promedio es menor al 50%
MKOOPKLA	59% vive en zonas donde el porcentaje de personas con poder adquisitivo es menor al 50%

Tabla 3.7 Características de los datos de entrenamiento (variables sociodemográficas).

En resumen, las principales características de las zonas donde viven los clientes es que son zonas de nivel social medio y bajo, con bajo nivel educativo, la religión más común es la protestante, hay familias con adultos, la mayoría son casados, la relación entre personas que rentan y personas con casa propia es alrededor de 50% - 50% y la mayoría cuenta con un auto.

3.2.1 Variables de números de pólizas (todos los datos)

Estas variables se refieren al número de pólizas que tienen los clientes. Existe una variable por cada tipo de póliza. Las variables pueden tener valores entre cero y doce.

Las siguientes dos tablas muestran las frecuencias de los distintos valores de las variables de números de pólizas. La primera tabla muestra las frecuencias absolutas por cada tipo de póliza. La segunda tabla muestra los porcentajes (frecuencias normalizadas) en lugar de las frecuencias absolutas.

Tipo Póliza	Frecuencias de los distintos valores de las variables de número de pólizas												
	0	1	2	3	4	5	6	7	8	9	10	11	12
AWAPART	3482	2334	6	0	0	0	0	0	0	0	0	0	0
AWABEDR	5740	81	0	0	0	1	0	0	0	0	0	0	0
AWALAND	5702	120	0	0	0	0	0	0	0	0	0	0	0
APERSAUT	2845	2712	246	12	5	0	1	1	0	0	0	0	0
ABESAUT	5774	40	4	3	1	0	0	0	0	0	0	0	0
AMOTSCO	5600	211	10	0	0	0	0	0	1	0	0	0	0
AVRAAUT	5813	6	2	1	0	0	0	0	0	0	0	0	0
AAANHANG	5757	59	4	2	0	0	0	0	0	0	0	0	0
ATRACTOR	5679	105	29	3	6	0	0	0	0	0	0	0	0
AWERKT	5801	12	6	2	0	0	1	0	0	0	0	0	0
ABROM	5426	382	14	0	0	0	0	0	0	0	0	0	0
ALEVEN	5529	173	100	11	8	0	0	0	1	0	0	0	0
APERSONG	5791	31	0	0	0	0	0	0	0	0	0	0	0
AGEZONG	5784	38	0	0	0	0	0	0	0	0	0	0	0
AWAOREG	5799	19	4	0	0	0	0	0	0	0	0	0	0
ABRAND	2666	3017	126	7	3	2	0	1	0	0	0	0	0
AZEILPL	5819	3	0	0	0	0	0	0	0	0	0	0	0
APLEZIER	5789	31	2	0	0	0	0	0	0	0	0	0	0
AFIETS	5675	111	34	2	0	0	0	0	0	0	0	0	0
AINBOED	5777	44	1	0	0	0	0	0	0	0	0	0	0
ABYSTAND	5740	81	1	0	0	0	0	0	0	0	0	0	0

Tabla 3.8 Frecuencias de los distintos valores de las variables de número de pólizas.

Tipo Póliza	Porcentajes de los distintos valores de las variables de número de pólizas												
	0	1	2	3	4	5	6	7	8	9	10	11	12
AWAPART	59.81	40.09	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AWABEDR	98.59	1.39	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AWALAND	97.94	2.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APERSAUT	48.87	46.58	4.23	0.21	0.09	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00
ABESAUT	99.18	0.69	0.07	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AMOTSCO	96.19	3.62	0.17	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
AVRAAUT	99.85	0.10	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AAANHANG	98.88	1.01	0.07	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ATRACTOR	97.54	1.80	0.50	0.05	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AWERKT	99.64	0.21	0.10	0.03	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
ABROM	93.20	6.56	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ALEVEN	94.97	2.97	1.72	0.19	0.14	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00

Tipo Póliza	Porcentajes de los distintos valores de las variables de número de pólizas												
	0	1	2	3	4	5	6	7	8	9	10	11	12
APERSONG	99.47	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGEZONG	99.35	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AWAOREG	99.60	0.33	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABRAND	45.79	51.82	2.16	0.12	0.05	0.03	0.00	0.02	0.00	0.00	0.00	0.00	0.00
AZEILPL	99.95	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APLEZIER	99.43	0.53	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFIETS	97.48	1.91	0.58	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AINBOED	99.23	0.76	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABYSTAND	98.59	1.39	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Tabla 3.9 Frecuencias normalizadas de los distintos valores de las variables de número de pólizas.

En las tablas de frecuencias de estas variables se observa que la mayoría de las variables presenta sólo dos valores. Las únicas variables que resaltan por tener mayor variación son:

AWAPART, (Number of private third party insurance)
APERSAUT (Number of car policies)
ABRAND (Number of fire policies)

Esto es más notorio en la gráfica de la tabla 3.9 que se muestra en la siguiente figura.

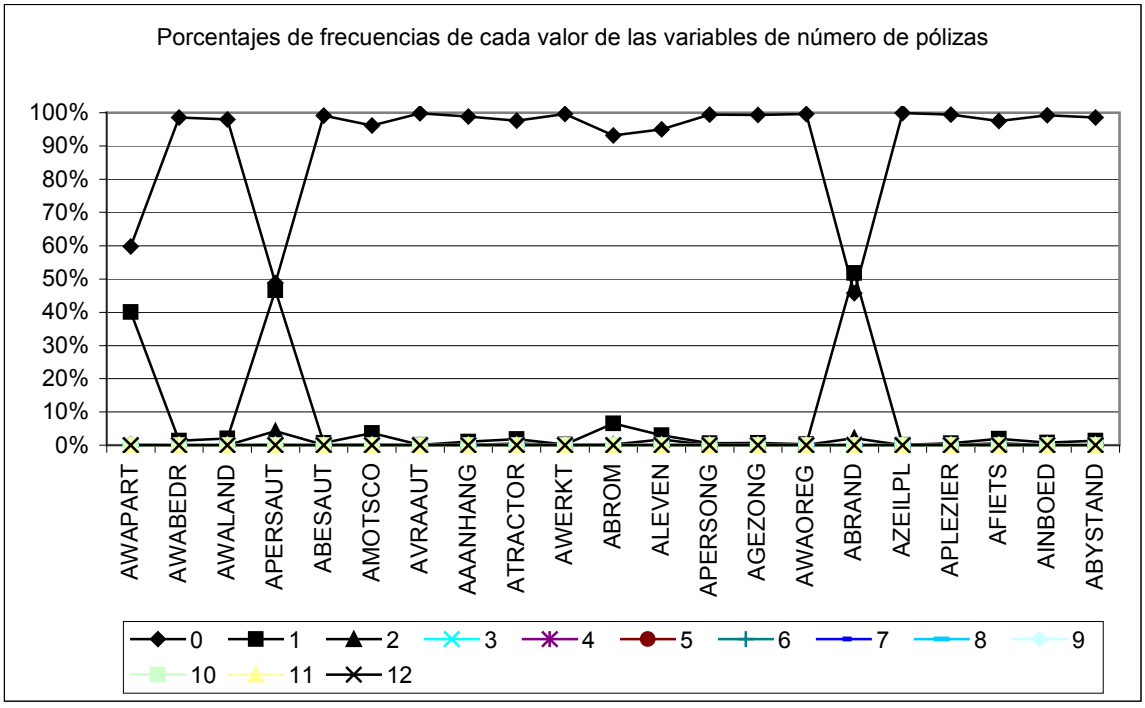

Fig. 3.3 Frecuencias normalizadas de cada valor de las variables de número de pólizas.

Variables de contribuciones de pólizas (todos los datos)

Estas variables se refieren a las contribuciones que hacen los clientes a las pólizas que tienen. Existe una variable por cada tipo de póliza. Las variables ya contienen datos

discretizados y pueden tener valores entre cero y nueve. Cada número corresponde a un rango de contribución. Entre mayor sea el número, mayor es la contribución. Un valor de cero indica que no hay contribución a esa póliza.

Las siguientes dos tablas muestran las frecuencias de los distintos valores de las variables de contribuciones de pólizas. La primera tabla muestra las frecuencias absolutas por cada tipo de póliza. La segunda tabla muestra los porcentajes (frecuencias normalizadas) en lugar de las frecuencias absolutas.

Tipo Póliza	Frecuencias de los distintos valores de las variables de contribuciones por tipo de póliza									
	0	1	2	3	4	5	6	7	8	9
PWAPART	3482	201	2128	11	0	0	0	0	0	0
PWABEDR	5740	7	30	23	17	1	4	0	0	0
PWALAND	5702	0	3	57	60	0	0	0	0	0
PPERSAUT	2845	0	0	0	1	613	2319	41	3	0
PBESAUT	5774	0	0	0	0	10	35	3	0	0
PMOTSCO	5600	0	0	3	136	32	49	2	0	0
PVRAAUT	5813	0	0	0	1	0	7	0	0	1
PAANHANG	5757	19	38	6	1	1	0	0	0	0
PTRACTOR	5679	0	0	79	27	28	9	0	0	0
PWERKT	5801	0	4	6	8	0	3	0	0	0
PBROM	5426	0	34	282	63	16	1	0	0	0
PLEVEN	5529	9	28	84	94	35	38	3	1	1
PPERSONG	5791	3	18	4	3	1	2	0	0	0
PGEZONG	5784	0	25	13	0	0	0	0	0	0
PWAOREG	5799	0	0	0	1	1	19	2	0	0
PBRAND	2666	161	535	920	1226	149	155	9	1	0
PZEILPL	5819	2	0	1	0	0	0	0	0	0
PPLEZIER	5789	5	5	5	13	2	3	0	0	0
PFIETS	5675	147	0	0	0	0	0	0	0	0
PINBOED	5777	18	16	6	3	1	1	0	0	0
PBYSTAND	5740	0	15	22	44	1	0	0	0	0

Tabla 3.10 Frecuencias de los distintos valores de las variables de contribuciones por tipo de póliza.

Tipo Póliza	Porcentajes de los distintos valores de las variables de contribuciones por tipo de póliza									
	0	1	2	3	4	5	6	7	8	9
PWAPART	59.81	3.45	36.55	0.19	0.00	0.00	0.00	0.00	0.00	0.00
PWABEDR	98.59	0.12	0.52	0.40	0.29	0.02	0.07	0.00	0.00	0.00
PWALAND	97.94	0.00	0.05	0.98	1.03	0.00	0.00	0.00	0.00	0.00
PPERSAUT	48.87	0.00	0.00	0.00	0.02	10.53	39.83	0.70	0.05	0.00
PBESAUT	99.18	0.00	0.00	0.00	0.00	0.17	0.60	0.05	0.00	0.00
PMOTSCO	96.19	0.00	0.00	0.05	2.34	0.55	0.84	0.03	0.00	0.00
PVRAAUT	99.85	0.00	0.00	0.00	0.02	0.00	0.12	0.00	0.00	0.02
PAANHANG	98.88	0.33	0.65	0.10	0.02	0.02	0.00	0.00	0.00	0.00
PTRACTOR	97.54	0.00	0.00	1.36	0.46	0.48	0.15	0.00	0.00	0.00
PWERKT	99.64	0.00	0.07	0.10	0.14	0.00	0.05	0.00	0.00	0.00
PBROM	93.20	0.00	0.58	4.84	1.08	0.27	0.02	0.00	0.00	0.00

Tipo Póliza	Porcentajes de los distintos valores de las variables de contribuciones por tipo de póliza									
	0	1	2	3	4	5	6	7	8	9
PLEVEN	94.97	0.15	0.48	1.44	1.61	0.60	0.65	0.05	0.02	0.02
PPERSONG	99.47	0.05	0.31	0.07	0.05	0.02	0.03	0.00	0.00	0.00
PGEZONG	99.35	0.00	0.43	0.22	0.00	0.00	0.00	0.00	0.00	0.00
PWAOREG	99.60	0.00	0.00	0.00	0.02	0.02	0.33	0.03	0.00	0.00
PBRAND	45.79	2.77	9.19	15.80	21.06	2.56	2.66	0.15	0.02	0.00
PZEILPL	99.95	0.03	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
PPLEZIER	99.43	0.09	0.09	0.09	0.22	0.03	0.05	0.00	0.00	0.00
PFIETS	97.48	2.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PINBOED	99.23	0.31	0.27	0.10	0.05	0.02	0.02	0.00	0.00	0.00
PBYSTAND	98.59	0.00	0.26	0.38	0.76	0.02	0.00	0.00	0.00	0.00

Tabla 3.11 Frecuencias normalizadas de los distintos valores de las variables de contribuciones por tipo de póliza.

Al igual que en las frecuencias de variables de número de pólizas, en los frecuencias de variables de contribuciones de pólizas se observa que hay poca variación en las variables, aunque ésta es mayor que la habida en las variables de número de pólizas. Las variables que resaltan por tener mayor variación corresponden a los mismos tipos de poliza:

PWAPART (Contribution private third party insurance)
PPERSAUT (Contribution car policies)
PBRAND (Contribution fire policies)

Nuevamente, esto es más notorio en la gráfica de la tabla 3.11 que se muestra en la siguiente figura:

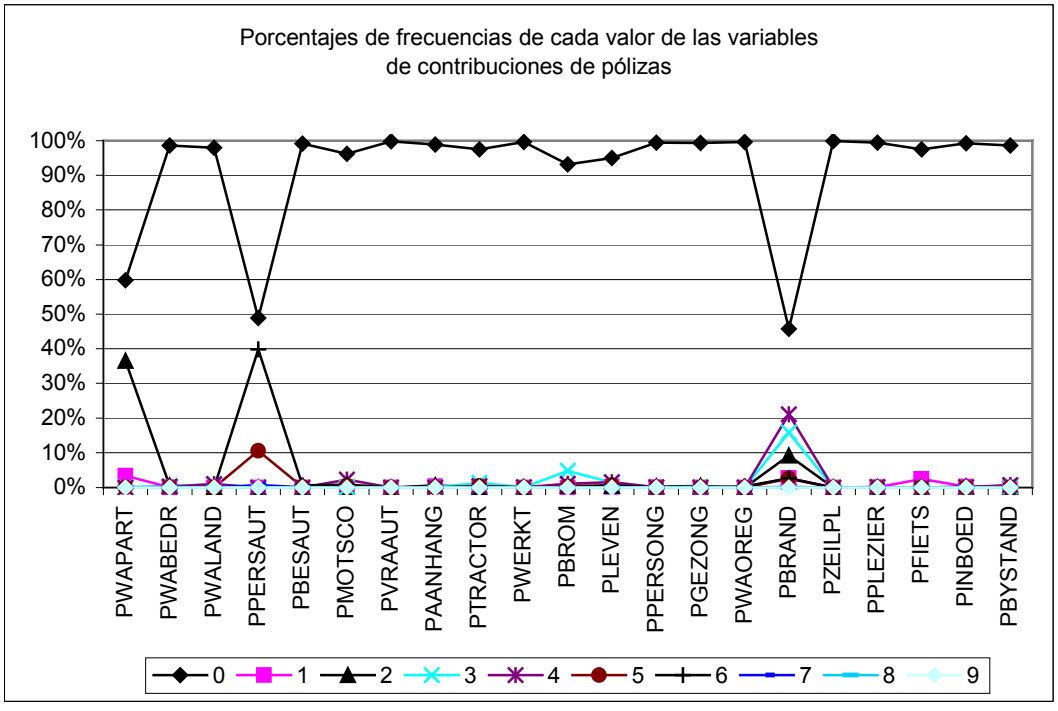

Fig. 3.4 Frecuencias normalizadas de cada valor de las variables de contribuciones de pólizas.

3.2.2 Variables sociodemográficas (datos de la clase)

Una manera de encontrar cuáles son las características que definen a los clientes que pertenecen a la clase (clientes que poseen póliza de seguro de casa rodante) es obtener nuevamente las tablas de frecuencias para este conjunto de datos y comparar las frecuencias para resaltar las diferencias.

Las siguientes cuatro tablas muestran las frecuencias de los distintos valores de las variables sociodemográficas considerando solamente los datos de la clase. Debido a que la variable MOSTYPE tiene 41 posibles valores, se muestran sus frecuencias en tablas separadas al resto de las 42 variables sociodemográficas, ya que estas últimas sólo tienen hasta 10 posibles valores.

Variable	Frecuencias de los distintos valores de MOSTYPE (datos de la clase)									
MOSTYPE	1	2	3	4	5	6	7	8	9	10
	13	6	25	2	2	12	3	51	12	9
	11	12	13	14	15	16	17	18	19	20
	9	16	13	0	0	0	0	0	0	2
	21	22	23	24	25	26	27	28	29	30
	0	4	4	5	2	1	1	0	2	4
	31	32	33	34	35	36	37	38	39	40
	6	8	46	9	8	16	10	23	19	0
	41									
	5									

Tabla 3.12 Frecuencias de los distintos valores de MOSTYPE (datos de la clase).

Variable	Frecuencias de los distintos valores de 42 variables sociodemográficas (datos de la clase)										
	0	1	2	3	4	5	6	7	8	9	10
MAANTHUI		315	33	0	0	0	0	0	0	0	0
MGEMOMV		8	115	171	50	4	0				
MGEMLEEF		1	87	183	64	12	1				
MOSHOOFD		48	66	59	0	15	4	20	89	42	5
MGODRK	177	107	54	7	1	1	1	0	0	0	
MGODPR	1	5	24	29	87	97	39	55	3	8	
MGODOV	130	93	99	18	7	1	0	0	0	0	
MGODGE	37	21	69	87	70	51	4	9	0	0	
MRELGE	2	1	3	6	10	48	71	116	25	66	
MRELSA	161	120	59	7	1	0	0	0	0	0	
MRELOV	93	33	110	74	26	6	4	1	0	1	
MFALLEEN	128	64	75	41	21	13	4	1	1	0	
MFGEKIND	23	24	57	89	88	30	27	8	1	1	
MFWEKIND	5	10	39	59	66	58	56	23	18	14	
MOPLHOOG	91	73	70	39	37	21	8	6	2	1	
MOPLMIDD	14	22	48	80	86	54	19	18	3	4	
MOPLLAAG	29	27	65	47	51	49	32	27	13	8	
MBERHOOG	73	61	78	53	33	13	18	14	3	2	

Variable	Frecuencias de los distintos valores de 42 variables sociodemográficas (datos de la clase)										
	0	1	2	3	4	5	6	7	8	9	10
MAANTHUI		315	33	0	0	0	0	0	0	0	0
MGEMOMV		8	115	171	50	4	0				
MBERZELF	233	82	26	3	1	3	0	0	0	0	
MBERBOER	284	36	20	6	1	1	0	0	0	0	
MBERMIDD	35	19	85	68	61	33	17	22	0	8	
MBERARBG	77	83	86	49	25	13	6	5	3	1	
MBERARBO	78	76	82	50	38	15	6	2	1	0	
MSKA	84	80	67	48	29	14	12	13	1	0	
MSKB1	72	81	101	65	19	4	5	0	1	0	
MSKB2	58	47	100	73	47	17	6	0	0	0	
MSKC	25	30	71	61	62	46	29	10	8	6	
MSKD	188	98	40	14	5	1	1	1	0	0	
MHHUUR	94	37	38	39	26	25	23	19	16	31	
MHKOOP	31	16	19	23	25	26	39	38	37	94	
MAUT1	0	0	1	7	13	59	91	119	19	39	
MAUT2	107	84	112	22	19	3	1	0	0	0	
MAUT0	121	48	108	49	13	4	5	0	0	0	
MZFONDS	7	0	27	15	28	72	45	87	28	39	
MZPART	39	28	87	45	72	28	15	27	0	7	
MINKM30	98	54	91	50	21	17	9	7	1	0	
MINK3045	26	18	55	74	84	45	25	10	2	9	
MINK4575	38	24	48	84	88	41	11	7	3	4	
MINK7512	152	97	58	19	14	7	0	0	0	1	
MINK123M	289	50	8	1	0	0	0	0	0	0	
MINKGEM	0	1	20	69	139	70	24	17	8	0	
MKOOPKLA	0	18	15	71	46	30	66	67	35	0	

Tabla 3.13 Frecuencias de los distintos valores de 42 variables sociodemográficas (datos de la clase).

Variable	Porcentajes de los distintos valores de MOSTYPE (datos de la clase)									
MOSTYPE	1	2	3	4	5	6	7	8	9	10
	3.74	1.72	7.18	0.57	0.57	3.45	0.86	14.66	3.45	2.59
	11	12	13	14	15	16	17	18	19	20
	2.59	4.60	3.74	0.00	0.00	0.00	0.00	0.00	0.00	0.57
	21	22	23	24	25	26	27	28	29	30
	0.00	1.15	1.15	1.44	0.57	0.29	0.29	0.00	0.57	1.15
	31	32	33	34	35	36	37	38	39	40
	1.72	2.30	13.22	2.59	2.30	4.60	2.87	6.61	5.46	0.00
	41									
	1.44									

Tabla 3.14 Frecuencias normalizadas de los distintos valores de MOSTYPE (datos de la clase).

Variable	Porcentajes de los distintos valores de 42 variables sociodemográficas (datos de la clase)										
	0	1	2	3	4	5	6	7	8	9	10
MAANTHUI		90.52	9.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MGEMOMV		2.30	33.05	49.14	14.37	1.15	0.00				
MGEMLEEF		0.29	25.00	52.59	18.39	3.45	0.29				
MOSHOOFD		13.79	18.97	16.95	0.00	4.31	1.15	5.75	25.57	12.07	1.44
MGODRK	50.86	30.75	15.52	2.01	0.29	0.29	0.29	0.00	0.00	0.00	
MGODPR	0.29	1.44	6.90	8.33	25.00	27.87	11.21	15.80	0.86	2.30	
MGODOV	37.36	26.72	28.45	5.17	2.01	0.29	0.00	0.00	0.00	0.00	
MGODGE	10.63	6.03	19.83	25.00	20.11	14.66	1.15	2.59	0.00	0.00	
MRELGE	0.57	0.29	0.86	1.72	2.87	13.79	20.40	33.33	7.18	18.97	
MRELSA	46.26	34.48	16.95	2.01	0.29	0.00	0.00	0.00	0.00	0.00	
MRELOV	26.72	9.48	31.61	21.26	7.47	1.72	1.15	0.29	0.00	0.29	
MFALLEEN	36.78	18.39	21.55	11.78	6.03	3.74	1.15	0.29	0.29	0.00	
MFGEKIND	6.61	6.90	16.38	25.57	25.29	8.62	7.76	2.30	0.29	0.29	
MFWEKIND	1.44	2.87	11.21	16.95	18.97	16.67	16.09	6.61	5.17	4.02	
MOPLHOOG	26.15	20.98	20.11	11.21	10.63	6.03	2.30	1.72	0.57	0.29	
MOPLMIDD	4.02	6.32	13.79	22.99	24.71	15.52	5.46	5.17	0.86	1.15	
MOPLLAAG	8.33	7.76	18.68	13.51	14.66	14.08	9.20	7.76	3.74	2.30	
MBERHOOG	20.98	17.53	22.41	15.23	9.48	3.74	5.17	4.02	0.86	0.57	
MBERZELF	66.95	23.56	7.47	0.86	0.29	0.86	0.00	0.00	0.00	0.00	
MBERBOER	81.61	10.34	5.75	1.72	0.29	0.29	0.00	0.00	0.00	0.00	
MBERMIDD	10.06	5.46	24.43	19.54	17.53	9.48	4.89	6.32	0.00	2.30	
MBERARBG	22.13	23.85	24.71	14.08	7.18	3.74	1.72	1.44	0.86	0.29	
MBERARBO	22.41	21.84	23.56	14.37	10.92	4.31	1.72	0.57	0.29	0.00	
MSKA	24.14	22.99	19.25	13.79	8.33	4.02	3.45	3.74	0.29	0.00	
MSKB1	20.69	23.28	29.02	18.68	5.46	1.15	1.44	0.00	0.29	0.00	
MSKB2	16.67	13.51	28.74	20.98	13.51	4.89	1.72	0.00	0.00	0.00	
MSKC	7.18	8.62	20.40	17.53	17.82	13.22	8.33	2.87	2.30	1.72	
MSKD	54.02	28.16	11.49	4.02	1.44	0.29	0.29	0.29	0.00	0.00	
MHHUUR	27.01	10.63	10.92	11.21	7.47	7.18	6.61	5.46	4.60	8.91	
MHKOOP	8.91	4.60	5.46	6.61	7.18	7.47	11.21	10.92	10.63	27.01	
MAUT1	0.00	0.00	0.29	2.01	3.74	16.95	26.15	34.20	5.46	11.21	
MAUT2	30.75	24.14	32.18	6.32	5.46	0.86	0.29	0.00	0.00	0.00	
MAUT0	34.77	13.79	31.03	14.08	3.74	1.15	1.44	0.00	0.00	0.00	
MZFONDS	2.01	0.00	7.76	4.31	8.05	20.69	12.93	25.00	8.05	11.21	
MZPART	11.21	8.05	25.00	12.93	20.69	8.05	4.31	7.76	0.00	2.01	
MINKM30	28.16	15.52	26.15	14.37	6.03	4.89	2.59	2.01	0.29	0.00	
MINK3045	7.47	5.17	15.80	21.26	24.14	12.93	7.18	2.87	0.57	2.59	
MINK4575	10.92	6.90	13.79	24.14	25.29	11.78	3.16	2.01	0.86	1.15	
MINK7512	43.68	27.87	16.67	5.46	4.02	2.01	0.00	0.00	0.00	0.29	
MINK123M	83.05	14.37	2.30	0.29	0.00	0.00	0.00	0.00	0.00	0.00	
MINKGEM	0.00	0.29	5.75	19.83	39.94	20.11	6.90	4.89	2.30	0.00	
MKOOPKLA	0.00	5.17	4.31	20.40	13.22	8.62	18.97	19.25	10.06	0.00	

Tabla 3.15 Frecuencias normalizadas de los distintos valores de 42 variables sociodemográficas (datos de la clase).

La siguiente tabla muestra las características de los datos sociodemográficos considerando sólamente los datos de los clientes que pertenecen a la clase.

Variable	Lo qué nos dice la variable
MOSTYPE	15% vive en zonas con familias de clase media (MOSTYPE = 8) 14% vive en zonas con familias grandes de clase baja (MOSTYPE = 33)
MAANTHUI	91% vive en zonas donde las personas tienen una casa (MAANTHUI = 1)
MGEMOMV	49% vive en zonas con 3 habitantes por casa (MGEMOMV = 3)
MGEMLEEF	53% vive en zonas donde la edad promedio está entre 40 y 50 años (MGEMLEEF = 3)
MOSHOOFD	26% vive en zonas que tienen familias con adultos (MOSHOOFD = 8)
MGODRK	99% vive en zonas donde el porcentaje de católicos es menor al 50% 51% vive en zonas donde el porcentaje de católicos es 0%
MGODPR	42% vive en zonas donde el porcentaje de protestantes es menor al 50%
MGODOV	99% vive en zonas donde el porcentaje de personas con otra religión es menor al 50% 37% vive en zonas donde el porcentaje de personas que no tienen otra religión es 0%
MGODGE	82% vive en zonas donde el porcentaje de personas sin religión es menor al 50%
MRELGE	6% vive en zonas donde el porcentaje de casados es menor al 50%
MRELSA	100% vive en zonas donde el porcentaje de personas en unión libre es menor al 50% 46% vive en zonas donde el porcentaje de personas en unión libre es 0%
MRELOV	97% vive en zonas donde el porcentaje de personas con otro tipo de relación es menor al 50%
MFALLEEN	95% vive en zonas donde el porcentaje de solteros es menor al 50%
MFGEKIND	81% vive en zonas donde el porcentaje de familias sin hijos es menor al 50%
MFWEKIND	51% vive en zonas donde el porcentaje de familias con hijos es menor al 55%
MOPLHOOG	89% vive en zonas donde el porcentaje de nivel educativo alto es menor al 50%
MOPLMIDD	72% vive en zonas donde el porcentaje de nivel educativo medio es menor al 50%
MOPLLAAG	63% vive en zonas donde el porcentaje de nivel educativo bajo es menor al 50%
MBERHOOG	86% vive en zonas donde el porcentaje de un estatus alto es menor al 50%
MBERZELF	99% vive en zonas donde el porcentaje de empresarios es menor al 50%
MBERBOER	100% vive en zonas donde el porcentaje de granjeros es menor al 50%
MBERMIDD	77% vive en zonas donde el porcentaje de administradores es menor al 50%
MBERARBG	92% vive en zonas donde el porcentaje de trabajadores capacitados es menor al 50%
MBERARBO	93% vive en zonas donde el porcentaje de trabajadores no capacitados es menor al 50%
MSKA	89% vive en zonas donde el porcentaje de personas de clase social A es menor al 50%
MSKB1	97% vive en zonas donde el porcentaje de personas de clase social B1 es menor al 50%
MSKB2	93% vive en zonas donde el porcentaje de personas de clase social B2 es menor al 50%
MSKC	72% vive en zonas donde el porcentaje de personas de clase social C es menor al 50%
MSKD	99% vive en zonas donde el porcentaje de personas de clase social D es menor al 50%
MHHUUR	67% vive en zonas donde el porcentaje de personas que rentan casa es menor al 50%

Variable	Lo qué nos dice la variable
MHKOOP	33% vive en zonas donde el porcentaje de personas con casa propia es menor al 50%
MAUT1	6% vive en zonas donde el porcentaje de personas con 1 auto es menor al 50%
MAUT2	99% vive en zonas donde el porcentaje de personas con 2 autos es menor al 50%
MAUT0	97% vive en zonas donde el porcentaje de personas sin auto es menor al 50%
MZFONDS	22% vive en zonas donde el porcentaje de personas con servicio nacional de salud es menor al 50%
MZPART	78% vive en zonas donde el porcentaje de personas con servicio privado de salud es menor al 50%
MINKM30	90% vive en zonas donde el porcentaje de personas con ingresos menores a 30,000 florines es menor al 50%
MINK3045	74% vive en zonas donde el porcentaje de personas con ingresos entre 30,000 y 45,000 florines es menor al 50%
MINK4575	81% vive en zonas donde el porcentaje de personas con ingresos entre 45,000 y 75,000 florines es menor al 50%
MINK7512	98% vive en zonas donde el porcentaje de personas con ingresos entre 75,000 y 122,000 florines es menor al 50%
MINK123M	100% vive en zonas donde el porcentaje de personas con ingresos mayores a 123,000 florines es menor al 50%
MINKGEM	67% vive en zonas donde el porcentaje de personas con ingresos promedio es menor al 50%
MKOOPKLA	43% vive en zonas donde el porcentaje de personas con poder adquisitivo es menor al 50%

Tabla 3.16 Características de los datos de entrenamiento (variables sociodemográficas).

Comparando estos porcentajes contra los porcentajes obtenidos al usar los datos de todos los clientes se observa que las diferencias van desde 1 punto porcentual hasta 16 puntos porcentuales. A continuación se listarán las variables para las cuales hubo una diferencia mayor o igual a 9 puntos porcentuales, así como la interpretación de las diferencias.

MRELGE: Del 15% que vive en zonas donde el porcentaje de casados es menor al 50% se pasó al 6%.

MOPLLAAG: Del 47% que vive en zonas donde el porcentaje de nivel educativo bajo es menor al 50% se pasó al 63%.

MHHUUR: Del 55% que vive en zonas donde el porcentaje de personas que rentan casa es menor al 50% se pasó al 67%.

MHKOOP: Del 45% que vive en zonas donde el porcentaje de personas con casa propia es menor al 50% se pasó al 33%.

MINKM30: Del 81% que vive en zonas donde el porcentaje de personas con ingresos menores a 30,000 florines es menor al 50% se pasó al 90%.

MINKGEM: Del 81% vive en zonas donde el porcentaje de personas con ingresos promedio es menor al 50% se pasó al 67%.

MKOOPKLA: Del 59% vive en zonas donde el porcentaje de personas con poder adquisitivo es menor al 50% se pasó al 43%.

La interpretación de estos porcentajes es que una persona es más probable de obtener una póliza si vive en una zona con:

Mayores porcentajes de casados.

Menores porcentajes de bajos niveles educativos.
Menores porcentajes de personas que rentan casas.
Mayores porcentajes de personas con casa propia.
Menores porcentajes de personas con ingresos menores a 30,000 florines.
Mayores porcentajes de personas con ingresos promedio.
Mayores porcentajes de personas con poder adquisitivo.

Aunque estos datos corresponden a zonas geográficas, no a individuos, de todas maneras nos dan una idea general del tipo de personas que adquieren una póliza de casa rodante: son personas casadas, con niveles educativos medios a altos, cuentan con casa propia, tienen ingresos mayores a 30,000 florines, ingresos mayores al promedio y tienen poder adquisitivo.

3.2.3 Variables de números de pólizas (datos de la clase)

Las siguientes dos tablas muestran las frecuencias de los distintos valores de las variables de número de pólizas considerando solamente el conjunto de clientes que pertenecen a la clase. La primera tabla muestra las frecuencias absolutas por cada tipo de póliza. La segunda tabla muestra los porcentajes (frecuencias normalizadas) en lugar de las frecuencias absolutas.

Tipo Póliza	Frecuencias de los distintos valores de las variables de número de pólizas (datos de la clase)												
	0	1	2	3	4	5	6	7	8	9	10	11	12
AWAPART	147	201	0	0	0	0	0	0	0	0	0	0	0
AWABEDR	343	5	0	0	0	0	0	0	0	0	0	0	0
AWALAND	345	3	0	0	0	0	0	0	0	0	0	0	0
APERSAUT	72	237	38	1	0	0	0	0	0	0	0	0	0
ABESAUT	346	2	0	0	0	0	0	0	0	0	0	0	0
AMOTSCO	332	15	1	0	0	0	0	0	0	0	0	0	0
AVRAAUT	348	0	0	0	0	0	0	0	0	0	0	0	0
AAANHANG	342	6	0	0	0	0	0	0	0	0	0	0	0
ATRACTOR	343	4	1	0	0	0	0	0	0	0	0	0	0
AWERKT	348	0	0	0	0	0	0	0	0	0	0	0	0
ABROM	340	8	0	0	0	0	0	0	0	0	0	0	0
ALEVEN	325	8	10	2	3	0	0	0	0	0	0	0	0
APERSONG	347	1	0	0	0	0	0	0	0	0	0	0	0
AGEZONG	342	6	0	0	0	0	0	0	0	0	0	0	0
AWAOREG	344	4	0	0	0	0	0	0	0	0	0	0	0
ABRAND	109	232	7	0	0	0	0	0	0	0	0	0	0
AZEILPL	347	1	0	0	0	0	0	0	0	0	0	0	0
APLEZIER	335	12	1	0	0	0	0	0	0	0	0	0	0
AFIETS	333	10	4	1	0	0	0	0	0	0	0	0	0
AINBOED	343	5	0	0	0	0	0	0	0	0	0	0	0
ABYSTAND	332	16	0	0	0	0	0	0	0	0	0	0	0

Tabla 3.17 Frecuencias de los distintos valores de las variables de números de pólizas (datos de la clase).

Tipo Póliza	Porcentajes de los distintos valores de las variables de número de pólizas (datos de la clase)												
	0	1	2	3	4	5	6	7	8	9	10	11	12
AWAPART	42.24	57.76	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AWABEDR	98.56	1.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AWALAND	99.14	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APERSAUT	20.69	68.10	10.92	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABESAUT	99.43	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AMOTSCO	95.40	4.31	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AVRAAUT	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AAANHANG	98.28	1.72	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ATRACTOR	98.56	1.15	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AWERKT	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABROM	97.70	2.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ALEVEN	93.39	2.30	2.87	0.57	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APERSONG	99.71	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AGEZONG	98.28	1.72	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AWAOREG	98.85	1.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABRAND	31.32	66.67	2.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AZEILPL	99.71	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APLEZIER	96.26	3.45	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFIETS	95.69	2.87	1.15	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AINBOED	98.56	1.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABYSTAND	95.40	4.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Tabla 3.18 Frecuencias normalizadas de los distintos valores de las variables de número de pólizas (datos de la clase).

Comparando la tabla de frecuencias normalizadas de los distintos valores de las variables de número de pólizas de todos los datos (tabla 3.9) contra la tabla que sólo considera los datos de clase (tabla 3.18) se encuentran las siguientes diferencias:

Los porcentajes de AWAPART = 0 cambiaron del 60% al 42%
Los porcentajes de AWAPART > 0 cambiaron del 40% al 58%
Los porcentajes de APERSAUT = 0 cambiaron del 49% al 21%
Los porcentajes de APERSAUT > 0 cambiaron del 51% al 79%
Los porcentajes de ABRAND = 0 cambiaron del 46% al 31%
Los porcentajes de ABRAND > 0 cambiaron del 54% al 69%

Estos números significan lo siguiente:

Las personas que están aseguradas contra daños a terceros tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas que tienen pólizas de auto tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas que tienen pólizas contra incendios tienen mayor probabilidad de obtener una póliza de casa rodante.

Quizás la razón de que las personas que cuenten con este tipo de pólizas tengan mayor probabilidad de obtener una casa rodante no se deba a la posesión de estos tipos de póliza, sino al hecho de que estas tres variables presentan una mayor varianza y por lo tanto son mejores discriminadoras de la clase.

La siguiente gráfica muestra la tabla de frecuencias normalizadas de los distintos valores de las variables de número de pólizas (tabla 3.18):

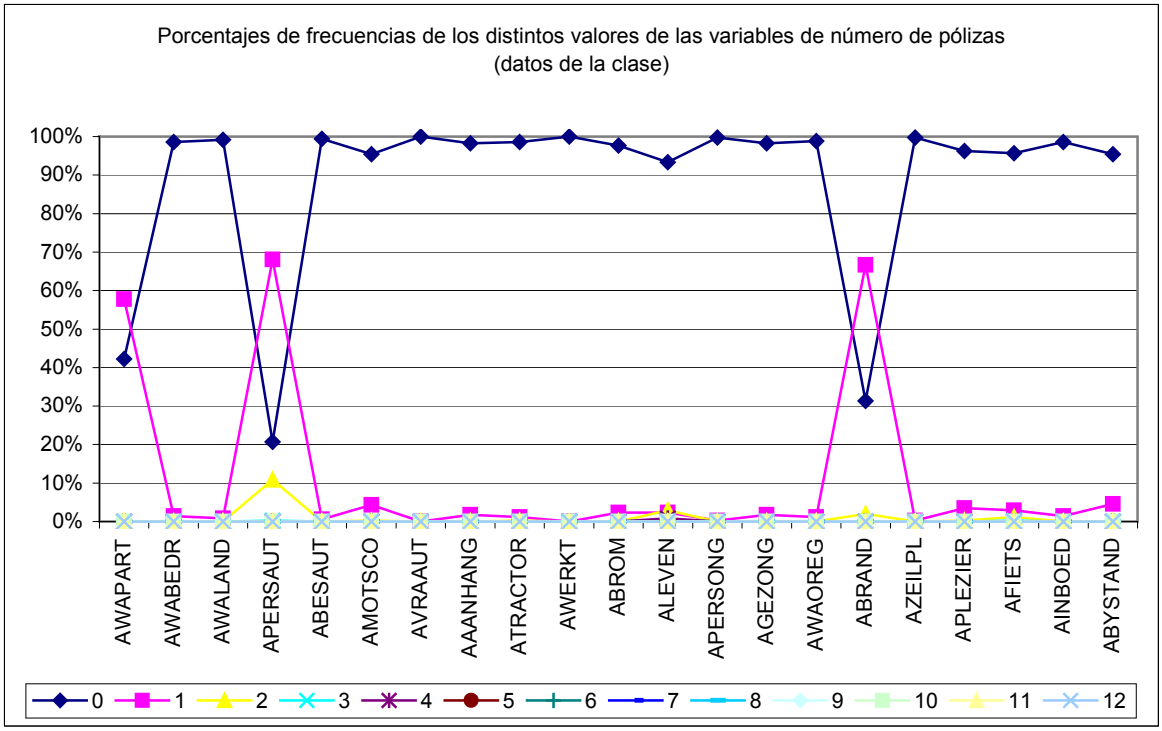

Fig. 3.5 Frecuencias normalizadas de los distintos valores de las variables de número de pólizas (datos de la clase).

Las diferencias mostradas arriba fueron con respecto al número de puntos porcentuales entre los porcentajes de todos los datos y porcentajes de los datos de la clase para una variable y valor particulares. También se pueden comparar los resultados tomando en cuenta la proporción del cambio. Considerando cambios mayores al 500% (es decir un aumento de cinco veces, o más) resaltan las siguientes variables:

Los porcentajes de ALEVEN = 4 cambiaron del 0.14% al 0.86%.
Los porcentajes de AZEILPL = 1 cambiaron del 0.05% al 0.29%.
Los porcentajes de APLEZIER = 1 cambiaron del 0.53% al 3.45%.
Los porcentajes de APLEZIER = 2 cambiaron del 0.03% al 0.29%.
Los porcentajes de AFIETS = 3 cambiaron del 0.03% al 0.29%.

Estos datos significan lo siguiente:

Las personas con varios seguros de vida tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas con un seguro de tabla de surfeo tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas con uno o dos seguros de bote tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas con tres pólizas de bicicleta tienen mayor probabilidad de obtener una póliza de casa rodante.

Cabe resaltar que aunque estos valores de variables se ven como buenos indicadores de la clase, tienen la desventaja de que su cobertura es muy pequeña y por lo tanto no son muy confiables estadísticamente. Sin embargo cuentan con buena predictibilidad y si se desea elegir un número pequeño de individuos con buena probabilidad de pertenecer a la clase, entonces se puede hacer usando estos indicadores.

3.2.4 Variables de contribuciones de pólizas (datos de la clase)

Las siguientes dos tablas muestran las frecuencias de los distintos valores de las variables de contribuciones de pólizas considerando solamente el conjunto de clientes que pertenecen a la clase. La primera tabla muestra las frecuencias absolutas por cada tipo de póliza. La segunda tabla muestra los porcentajes (frecuencias normalizadas) en lugar de las frecuencias absolutas.

Tipo Póliza	Frecuencias de los distintos valores de las variables de contribuciones por tipo de póliza (datos de la clase)									
	0	1	2	3	4	5	6	7	8	9
PWAPART	147	8	191	2	0	0	0	0	0	0
PWABEDR	343	0	2	3	0	0	0	0	0	0
PWALAND	345	0	0	2	1	0	0	0	0	0
PPERSAUT	72	0	0	0	0	14	262	0	0	0
-PBESAUT	346	0	0	0	0	0	2	0	0	0
PMOTSCO	332	0	0	2	9	4	1	0	0	0
PVRAAUT	348	0	0	0	0	0	0	0	0	0
PAANHANG	342	1	5	0	0	0	0	0	0	0
PTRACTOR	343	0	0	2	0	2	1	0	0	0
PWERKT	348	0	0	0	0	0	0	0	0	0
PBROM	340	0	1	6	0	1	0	0	0	0
PLEVEN	325	0	0	6	11	4	2	0	0	0
PPERSONG	347	0	1	0	0	0	0	0	0	0
PGEZONG	342	0	2	4	0	0	0	0	0	0
PWAOREG	344	0	0	0	0	0	4	0	0	0
PBRAND	109	3	6	68	151	8	3	0	0	0
PZEILPL	347	1	0	0	0	0	0	0	0	0
PPLEZIER	335	3	2	2	4	0	2	0	0	0
PFIETS	333	15	0	0	0	0	0	0	0	0
PINBOED	343	3	2	0	0	0	0	0	0	0
PBYSTAND	332	0	4	4	8	0	0	0	0	0

Tabla 3.19 Frecuencias de los distintos valores de las variables de contribuciones de pólizas (datos de la clase).

Tipo Póliza	Porcentajes de los distintos valores de las variables de contribuciones por tipo de póliza (datos de la clase)									
	0	1	2	3	4	5	6	7	8	9
PWAPART	42.24	2.30	54.89	0.57	0.00	0.00	0.00	0.00	0.00	0.00
PWABEDR	98.56	0.00	0.57	0.86	0.00	0.00	0.00	0.00	0.00	0.00
PWALAND	99.14	0.00	0.00	0.57	0.29	0.00	0.00	0.00	0.00	0.00
PPERSAUT	20.69	0.00	0.00	0.00	0.00	4.02	75.29	0.00	0.00	0.00
PBESAUT	99.43	0.00	0.00	0.00	0.00	0.00	0.57	0.00	0.00	0.00
PMOTSCO	95.40	0.00	0.00	0.57	2.59	1.15	0.29	0.00	0.00	0.00
PVRAAUT	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PAANHANG	98.28	0.29	1.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PTRACTOR	98.56	0.00	0.00	0.57	0.00	0.57	0.29	0.00	0.00	0.00
PWERKT	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PBROM	97.70	0.00	0.29	1.72	0.00	0.29	0.00	0.00	0.00	0.00
PLEVEN	93.39	0.00	0.00	1.72	3.16	1.15	0.57	0.00	0.00	0.00
PPERSONG	99.71	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PGEZONG	98.28	0.00	0.57	1.15	0.00	0.00	0.00	0.00	0.00	0.00
PWAOREG	98.85	0.00	0.00	0.00	0.00	0.00	1.15	0.00	0.00	0.00
PBRAND	31.32	0.86	1.72	19.54	43.39	2.30	0.86	0.00	0.00	0.00
PZEILPL	99.71	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PPLEZIER	96.26	0.86	0.57	0.57	1.15	0.00	0.57	0.00	0.00	0.00
PFIETS	95.69	4.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PINBOED	98.56	0.86	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PBYSTAND	95.40	0.00	1.15	1.15	2.30	0.00	0.00	0.00	0.00	0.00

Tabla 3.20 Frecuencias normalizadas de los distintos valores de las variables de contribuciones de pólizas (datos de la clase).

Comparando la tabla de frecuencias normalizadas de los distintos valores de las variables de contribuciones de pólizas de todos los datos (tabla 3.11) contra la tabla que sólo considera los datos de clase (tabla 3.20) se encuentran las siguientes diferencias:

Los porcentajes de PWAPART = 0 cambiaron del 60% Al 42%
Los porcentajes de PWAPART > 0 cambiaron del 40% Al 58%
Los porcentajes de PPERSAUT = 0 cambiaron del 49% al 21%
Los porcentajes de PPERSAUT > 0 cambiaron del 51% al 79%
Los porcentajes de PBRAND = 0 cambiaron del 46% al 31%
Los porcentajes de PBRAND > 0 cambiaron del 54% al 69%

Estos números significan lo mismo que en el caso de las variables de números de pólizas:

Las personas que están aseguradas contra daños a terceros tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas que tienen pólizas de auto tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas que tienen pólizas contra incendios tienen mayor probabilidad de obtener una póliza de casa rodante.

La siguiente gráfica muestra la tabla de frecuencias normalizadas de los distintos valores de las variables de contribuciones de pólizas (tabla 3.20):

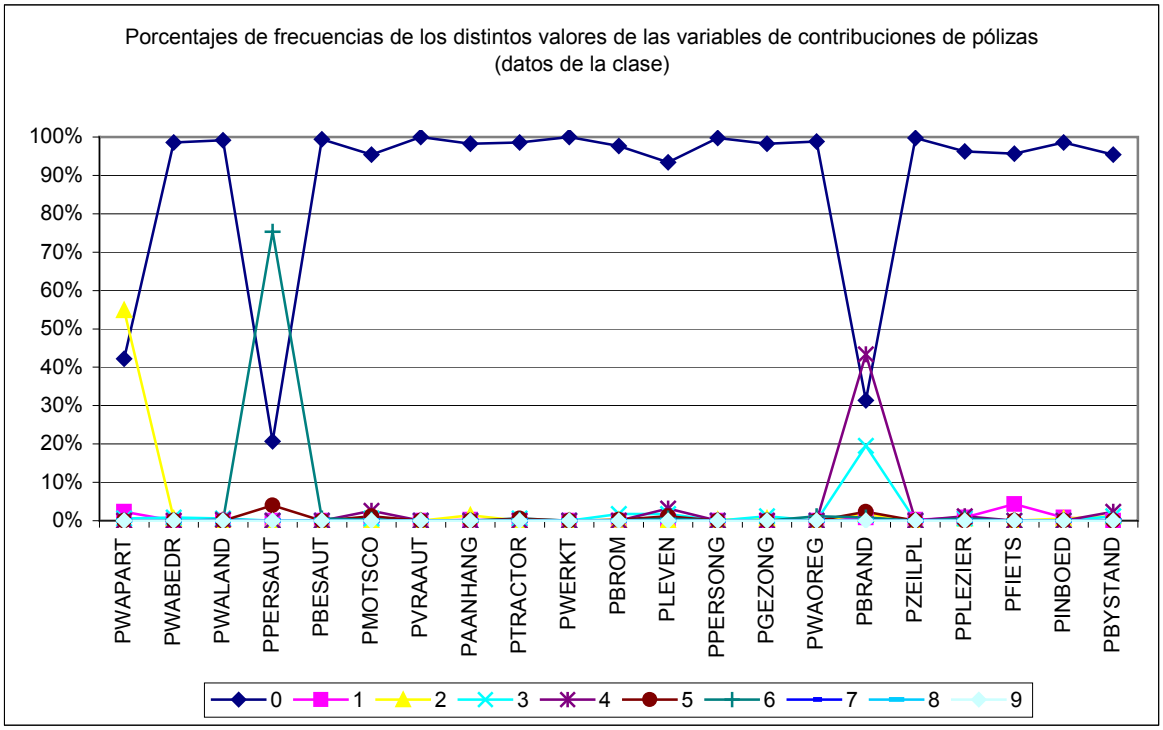

Fig. 3.6 Frecuencias normalizadas de los distintos valores de las variables de contribuciones de póliza (datos de la clase).

Tomando en cuenta la proporción del cambio (mayor a 500%) resaltan las siguientes variables:

Los porcentajes de PGEZONG = 3 cambiaron del 0.22% al 1.15%.
Los porcentajes de PZEILPL = 2 cambiaron del 0.03% al 0.29%.
Los porcentajes de PPLEZIER = 1 cambiaron del 0.09% al 0.86%.
Los porcentajes de PPLEZIER = 2 cambiaron del 0.09% al 0.57%.
Los porcentajes de PPLEZIER = 3 cambiaron del 0.09% al 0.57%.
Los porcentajes de PPLEZIER = 4 cambiaron del 0.22% al 1.15%.
Los porcentajes de PPLEZIER = 6 cambiaron del 0.05% al 0.57%.

Estos datos significan lo siguiente:

Las personas con una contribución mediana de seguros contra accidentes familiares tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas con una contribución baja de seguro de tabla de surfeo tienen mayor probabilidad de obtener una póliza de casa rodante.

Las personas con contribuciones mayores a cero de seguros de bote tienen mayor probabilidad de obtener una póliza de casa rodante.

Nuevamente, cabe resaltar que aunque estos valores de variables se ven como buenos indicadores de la clase, tienen la desventaja de que su cobertura es muy pequeña y por lo

tanto no son muy confiables estadísticamente. Sin embargo, es de interés notar que parecen ser buenos indicadores de la clase.

3.2.5 Funciones ε' y ε

Los resultados de ε' se muestran en la siguiente tabla.

Estos resultados se obtuvieron a partir de los datos de entrenamiento, el cual consiste en 5822 registros y 86 variables (incluyendo la variable objetivo).

Variable	Descripción	ε'
PPERSAUT	Contribución de pólizas de auto	13.7055
APERSAUT	Número de pólizas de auto	11.7175
PBRAND	Contribución de pólizas contra incendio	7.4781
MKOOPKLA	Clase con poder adquisitivo	7.3118
PWAPART	Contribución de seguros contra daño a terceros	7.1496
MINKGEM	Ingreso promedio	6.9940
AWAPART	Número de seguros contra daño a terceros	6.7954
MRELGE	Casado	6.2178
MHKOOP	Casa propia	6.1185
MAUT1	1 auto	5.9669
MOPLHOOG	Educación de nivel alto	5.6529
ABRAND	Número de pólizas contra incendio	5.2396
MINK4575	Ingresos 45 - 75,000	4.6293
MSKA	Clase social A	4.4096
MBERHOOG	Estatus alto	4.3945
MZPART	Segura privado de salud	4.1420
MINK7512	Ingresos 75 - 122,000	3.9779
MBERMIDD	Administración media	3.3065
MOPLMIDD	Educación de nivel medio	3.2968
APLEZIER	Número de pólizas de bote	3.2093
ABYSTAND	Número de pólizas de seguridad social	2.9741
PPLEZIER	Contribución de pólizas de bote	2.8791
PBYSTAND	Contribución de pólizas de seguridad social	2.8418
MGEMOMV	Habitantes promedio por casa	2.8317
MGODPR	Protestante	2.6197
MFWEKIND	Familia sin hijos	2.4609
MSKB1	Clase social B1	2.1502
ALEVEN	Número de seguros de vida	1.9512
AFIETS	Número de pólizas de bicicleta	1.7940
PFIETS	Contribución de pólizas de bicicleta	1.7111
PGEZONG	Contribución de pólizas de seguro contra accidentes familiares	1.7108
AGEZONG	Número de pólizas de seguro contra accidentes familiares	1.6133
MBERZELF	Empresario	1.6055
PLEVEN	Contribución de seguros de vida	1.4027
PWAOREG	Contribución de pólizas de seguro contra discapacidades	1.3942

Variable	Descripción	ε'
AWAOREG	Número de pólizas de seguro contra discapacidades	1.2548
AINBOED	Número de pólizas de seguro de propiedad	1.0583
AZEILPL	Número de pólizas de tablas de surfeo	0.8693
PAANHANG	Contribución de pólizas de tráiler	0.8467
PZEILPL	Contribución de pólizas de tablas de surfeo	0.7311
AAANHANG	Número de pólizas de tráiler	0.6956
AMOTSCO	Número de pólizas de motocicleta	0.6557
MFGEKIND	Familia con hijos	0.5905
MAUT2	2 autos	0.5898
MGODRK	Católico romano	0.5246
PINBOED	Contribución de pólizas de seguro de propiedad	0.4826
PMOTSCO	Contribución de pólizas de motocicleta	0.4257
MGEMLEEF	Edad promedio	0.3561
MSKB2	Clase social B2	0.3157
MGODOV	Otra religión	0.3011
AWABEDR	Número de seguros contra daño a terceros (empresas)	-0.0646
MINK3045	Ingresos 30 - 45,000	-0.0734
PWABEDR	Contribución de seguros contra daño a terceros (empresas)	-0.1609
MINK123M	Ingresos > 123,000	-0.1774
PBESAUT	Contribución de pólizas de camioneta repartidora	-0.5771
APERSONG	Número de pólizas de seguro contra accidentes	-0.8570
MAANTHUI	Número de casas	-1.0064
PTRACTOR	Contribución de pólizas de tractor	-1.0313
ABESAUT	Número de pólizas de camioneta repartidora	-1.1340
PPERSONG	Contribución de pólizas de seguro contra accidentes	-1.3213
ATRACTOR	Número de pólizas de tractor	-1.9986
AWALAND	Número de seguros contra daño a terceros (agricultura)	-2.3909
PWALAND	Contribución de seguros contra daño a terceros (agricultura)	-2.5241
AVRAAUT	Número de pólizas de camión de carga	-2.7123
MRELSA	Unión libre	-2.9454
PVRAAUT	Contribución de pólizas de camión de carga	-2.9461
MSKC	Clase social C	-3.0833
MGODGE	Sin religión	-3.0946
MBERARBG	Trabajadores capacitados	-3.2517
AWERKT	Número de pólizas de máquinas para agricultura	-3.7994
MZFONDS	Servicio nacional de salud	-4.1975
PWERKT	Contribución de pólizas de máquinas para agricultura	-4.3517
MBERARBO	Trabajadores no capacitados	-4.3544
MFALLEEN	Soltero	-4.5906
PBROM	Contribución de pólizas de ciclomotor	-5.3919
MRELOV	Otra relación	-5.4426
ABROM	Número de pólizas de ciclomotor	-5.7089
MBERBOER	Granjero	-5.8298
MSKD	Clase social D	-5.8722

Variable	Descripción	ε'
MHHUUR	Casa rentada	-6.1732
MAUT0	Sin auto	-6.7325
MOPLLAAG	Educación de nivel bajo	-6.9341
MINKM30	Ingresos < 30,000	-7.0924

Tabla 3.21 Valores de ε' (datos de entrenamiento).

Los valores más grandes de ε' indican las características de las personas más probables de obtener una póliza de casa rodante:

Personas con alto poder adquisitivo, casadas, poseen casa propia, un auto, cuentan con pólizas de auto, pólizas contra incendios y pólizas contra daño a terceros.

Por otra parte, los valores más pequeños de ε' indican las características de las personas menos probables de obtener una casa rodante, como son las personas con bajo poder adquisitivo, solteras, rentan casa, no poseen auto y cuentan con pólizas de ciclomotor.

A continuación se muestran los 25 valores más altos y los 25 valores mas bajos de ε: La lista total de valores de ε se muestra en el apéndice A. La columna $p(C|\boldsymbol{X})$ se calcula como 100 * N(C^$\boldsymbol{X}$) / N($\boldsymbol{X}$). La columna $p(\boldsymbol{X}|C)$ se calcula como 100 * N(C^$\boldsymbol{X}$) / N(**C**).

Variable	Valor	Descripción del valor	ε	N(C^$\boldsymbol{X}$)	N($\boldsymbol{X}$)	N(C)	$p(C\|\boldsymbol{X})$	$p(\boldsymbol{X}\|C)$
PPERSAUT	6	Contribución de pólizas de auto = f 1,000 – 4,999	10.8080	262	2319	348	11.30	75.29
PBRAND	4	Contribución de pólizas contra incendio = f 200 – 499	9.3628	151	1226	348	12.32	43.39
APLEZIER	1	Número de pólizas de bote = 1	7.6876	12	31	348	38.71	3.45
MKOOPKLA	7	Clase con poder adquisitivo = 76 - 88%	7.4918	67	474	348	14.14	19.25
MOSTYPE	8	Subtipo de cliente = Familias clase media	7.0419	51	339	348	15.04	14.66
MOSHOOFD	2	Tipo principal de cliente = Cultivadores	6.7765	66	502	348	13.15	18.97
APERSAUT	2	Número de pólizas de auto = 2	6.2653	38	246	348	15.45	10.92
APERSAUT	1	Número de pólizas de auto = 1	6.0665	237	2712	348	8.74	68.10
PWAPART	2	Contribución de seguros contra daño a terceros = f 50 – 99	5.8342	191	2128	348	8.98	54.89
AWAPART	1	Número de seguros contra daño a terceros = 1	5.3688	201	2334	348	8.61	57.76
ABYSTAND	1	Número de pólizas de seguridad social = 1	5.2298	16	81	348	19.75	4.60
MHHUUR	0	Casa rentada = 0%	5.1041	94	949	348	9.91	27.01
MHKOOP	9	Casa propia = 100%	5.1041	94	949	348	9.91	27.01
PPLEZIER	1	Contribución de pólizas de bote = f 1 – 49	5.0956	3	5	348	60.00	0.86
PMOTSCO	3	Contribución de pólizas de motocicleta = f 100 – 199	4.4341	2	3	348	66.67	0.57
PPLEZIER	6	Contribución de pólizas de bote = f 1,000 – 4,999	4.4341	2	3	348	66.67	0.57
MOPLLAAG	2	Educación de nivel bajo = 11 - 23%	4.1047	65	667	348	9.75	18.68
MOPLHOOG	4	Educación de nivel alto = 37 - 49%	4.0917	37	326	348	11.35	10.63
MINKGEM	5	Ingreso promedio = 50 - 62%	4.0799	70	733	348	9.55	20.11

Variable	Valor	Descripción del valor	ε	N(*C*^***X***)	N(***X***)	N(*C*)	$p(C|\boldsymbol{X})$	$p(\boldsymbol{X}|C)$
ABRAND	1	Número de pólizas contra incendio = 1	3.9676	232	3017	348	7.69	66.67
MSKA	7	Clase social A = 76 - 88%	3.9286	13	79	348	16.46	3.74
MBERARBG	1	Trabajadores capacitados = 1 - 10%	3.8848	83	921	348	9.01	23.85
MAUT1	7	1 auto = 76 - 88%	3.8760	119	1413	348	8.42	34.20
MAUT0	0	Sin auto = 0%	3.8028	121	1450	348	8.34	34.77
MGODPR	7	Protestante = 76 - 88%	3.7811	55	564	348	9.75	15.80
PBROM	3	Contribución de pólizas de ciclomotor = f 100 – 199	-2.7270	6	282	348	2.13	1.72
MAUT1	4	1 auto = 37 - 49%	-2.7459	13	448	348	2.90	3.74
MOPLLAAG	6	Educación de nivel bajo = 63 - 75%	-2.7633	32	856	348	3.74	9.20
MHKOOP	1	Casa propia = 1 - 10%	-2.8730	16	530	348	3.02	4.60
MHHUUR	8	Casa rentada = 89 - 99%	-2.8894	16	532	348	3.01	4.60
MOSHOOFD	10	Tipo principal de cliente = Granjeros	-2.9193	5	276	348	1.81	1.44
MOSTYPE	23	Subtipo de cliente = Jóvenes crecientes	-2.9296	4	251	348	1.59	1.15
MSKC	5	Clase social C = 50 - 62%	-2.9394	46	1168	348	3.94	13.22
MKOOPKLA	1	Clase con poder adquisitivo = 1 - 10%	-2.9749	18	587	348	3.07	5.17
MINKM30	5	Ingresos < 30,000 = 50 - 62%	-3.0002	17	568	348	2.99	4.89
MINK7512	0	Ingresos 75 - 122,000 = 0%	-3.1114	152	3246	348	4.68	43.68
MINKGEM	2	Ingreso promedio = 11 - 23%	-3.1267	20	651	348	3.07	5.75
ABROM	1	Número de pólizas de ciclomotor = 1	-3.2014	8	382	348	2.09	2.30
MOSHOOFD	5	Tipo principal de cliente = Buen nivel de vida	-3.3619	15	569	348	2.64	4.31
MOPLHOOG	0	Educación de nivel alto = 0%	-3.3987	91	2147	348	4.24	26.15
MAUT0	4	Sin auto = 37 - 49%	-3.8454	13	587	348	2.21	3.74
PPERSAUT	5	Contribución de pólizas de auto = f 500 – 999	-3.8574	14	613	348	2.28	4.02
PBRAND	0	Contribución de pólizas contra incendio = f 0	-4.1138	109	2666	348	4.09	31.32
ABRAND	0	Número de pólizas contra incendio = 0	-4.1138	109	2666	348	4.09	31.32
PWAPART	0	Contribución de seguros contra daño a terceros = f 0	-4.3699	147	3482	348	4.22	42.24
AWAPART	0	Número de seguros contra daño a terceros = 0	-4.3699	147	3482	348	4.22	42.24
MINKGEM	3	Ingreso promedio = 24 - 36%	-4.4608	69	1932	348	3.57	19.83
PBRAND	2	Contribución de pólizas contra incendio = f 50 – 99	-4.7377	6	535	348	1.12	1.72
PPERSAUT	0	Contribución de pólizas de auto = f 0	-7.7546	72	2845	348	2.53	20.69
APERSAUT	0	Número de pólizas de auto = 0	-7.7546	72	2845	348	2.53	20.69

Tabla 3.22 Principales valores de ε (datos de entrenamiento).

N(*C*^***X***) es el número de individuos que cumplen con el indicador ***X*** y que pertenecen a la clase, donde ***X*** se define como el par *variable* = *valor*.

N(***X***) es el número de individuos que cumplen con el indicador ***X***.

N(*C*) es el número de individuos que cumplen con la clase.

$p(C|\boldsymbol{X})$ es la probabilidad de la clase dado el indicador X.

$p(\boldsymbol{X}|C)$ es la probabilidad de tener el indicador X dada la clase.

Tanto ε como ε' indican cuáles son las características que tienen los individuos más probables de pertenecer a cierta clase dada, en este caso la clase de las personas que compran una casa rodante. La ventaja del uso de ε y ε' es que son medidas fáciles de interpretar, ya que están asociados directamente a las variables y a sus posibles valores.

En la tabla se observa que los valores de ε son consistentes con los de ε' y además tienen la ventaja de proporcionar una vista más detallada de las características de las personas con mayor certeza estadística de comprar una póliza. Por ejemplo, con ε' se supo que la variable "Contribución de pólizas de auto" es importante para determinar los individuos que pertenecen a la clase y con ε se obtuvo la información adicional de que el valor de esta contribución está principalmente entre 1,000 y 5,000 florines en los individuos que pertenecen a la clase.

Los indicadores de ε también pueden ser usados para realizar predicciones de la siguiente manera:

1. Se toma el indicador con el valor más alto de ε.
2. Se eligen todos los individuos que cumplen con este indicador hasta completar el número o porcentaje deseado de individuos.
3. Si no se completó el número de individuos deseados, se toma el siguiente indicador más alto y se regresa al paso 2 hasta que se haya completado el número deseado de individuos.

Otro empleo de ε y ε' es para seleccionar las variables que sean más convenientes para la predicción. Las variables con los valores más altos de ε y ε' son las que sirven para diferenciar mejor los individuos que pertenecen a la clase. Por el contrario, las variables con los valores más bajos sirven para diferenciar mejor las personas que no pertenecen a la clase.

Una ventaja importante del empleo de ε y ε' es que son sencillos de calcular y la complejidad del cálculo es lineal.

3.3 Predicción

El problema de predicción en los datos de CoIL consiste en dar una lista de quienes serían las personas interesadas en comprar una póliza de seguro para una casa rodante. El número de personas deseadas en la lista es el veinte por ciento (800 personas) del conjunto total.

Como se trata de un conjunto de datos pequeño y con relativamente pocas variables lo más recomendado es elegir un modelo sencillo de predicción. La clasificación Bayesiana ingenua es un modelo sencillo que en la práctica ha mostrado tener un buen desempeño de predicción.

3.3.1 Criterios para la selección de las variables

Un primer paso para realizar la predicción consiste en realizar la selección de las variables más adecuadas para realizar la predicción.

Tratándose de relativamente pocas variables (menos de 100) y de pocos datos (5822 en el conjunto de entrenamiento y 4000 en el conjunto de prueba) se pueden hacer varias pruebas para determinar qué subconjunto de variables funciona mejor para entrenar el modelo elegido y posteriormente realizar la predicción con dicho subconjunto.

Se pueden seguir varios criterios para elegir las variables más convenientes:

- Elegir las variables con mayores valores de probabilidad $p(C|\boldsymbol{X})$, donde $\boldsymbol{X}$ está compuesta por una sola variable, es decir $\boldsymbol{X}$ es de la forma *variable*=*valor*. La probabilidad $p(C|\boldsymbol{X})$ tiene la desventaja de que se pueden obtener valores muy altos, pero que ocurren para muy pocos casos de $\boldsymbol{X}$. Por ejemplo, si $\boldsymbol{X}$ = "PMOSTCO = 3", se tiene una probabilidad muy alta (66.67%), pero sólo hay tres datos cumplen con $\boldsymbol{X}$, (de los cuales 2 pertenecen a la clase). El criterio $p(C|\boldsymbol{X})$ es más conveniente cuando se desea seleccionar un número pequeño de individuos con mayor probabilidad de pertenecer a la clase, por ejemplo, los casos en donde se desee seleccionar el 1% de mejores individuos.

- Elegir las variables que tienen los valores más altos de ε'. ε' es una medida que sólo aplica para las variables métricas, por lo que no se puede calcular ε' para las variables MOSTYPE y MOSHOOFD.

- Elegir las variables con mayores valores de ε. Esta medida no tiene la restricción de que las variables deben ser métricas. Además, la ventaja de ε es que sus valores se normalizan por el número de datos que cumplen con X, por lo que se reduce el problema de tener valores altos de ε cuando X se cumple para pocos casos.

En la siguiente sección se analizarán los resultados de la clasificación de los datos de entrenamiento usando los distintos criterios de selección de variables.

3.3.2 Clasificación del conjunto de entrenamiento mediante clasificación Bayesiana ingenua

Criterio ε

La figura 3.7 muestra el desempeño de la clasificación, sobre los datos de entrenamiento, utilizando distintos números de variables, desde una variable, hasta 85 variables. Para la clasificación usando solamente una variable se eligió la variable con el valor más alto de ε de todos sus posibles valores. Para la clasificación usando dos variables se eligieron las dos variables con los valores más altos de $p(C|\boldsymbol{X})$ de todos sus posibles valores y así sucesivamente hasta llegar a las 85 variables. Esta es una manera práctica de probar cuáles variables son predictivas, ya que resulta imposible probar todas las combinaciones de las variables. El desempeño se mide como el número de aciertos en el 20% superior.

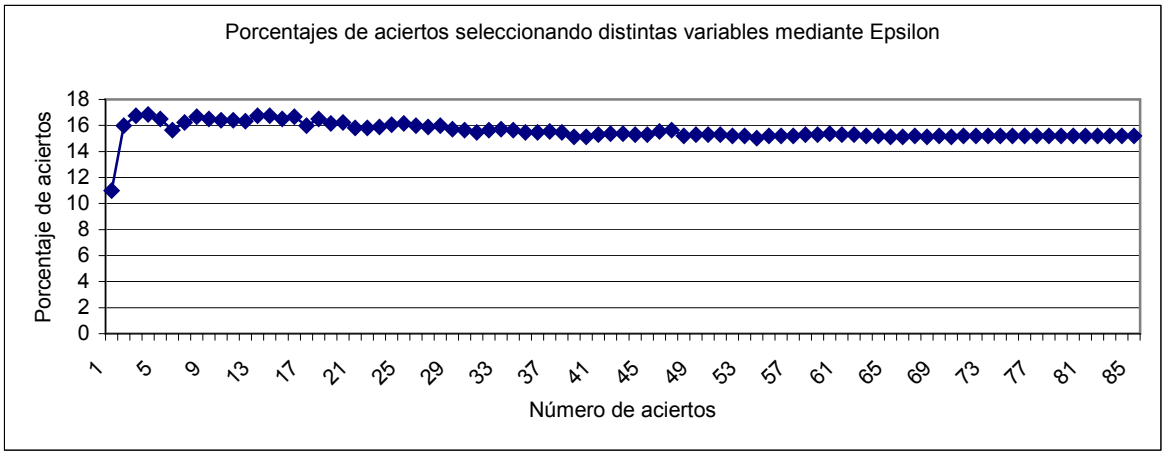

Fig. 3.7 Porcentaje de aciertos en el 20% superior seleccionando distintas variables mediante ε.

En la gráfica se observa que una sola variable (PPERSAUT) fue suficiente para proporcionar un desempeño mejor que el azar. Y el uso de solamente dos variables (PPERSAUT y PBRAND) proporcionó un muy buen desempeño.

También es interesante notar que el uso de más variables no necesariamente implica mejores desempeños en la clasificación. En la figura se aprecia que a partir de aproximadamente la variable número 30 los desempeños disminuyen ligeramente.

La siguiente tabla muestra los porcentajes de aciertos para los conjuntos de 1 a 20 variables.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
11.00	15.98	16.75	16.84	16.49	15.64	16.24	16.67	16.49	16.41	16.41	16.32	16.75	16.75	16.49	16.67	15.98	16.49	16.15	16.24

Tabla 3.23 Porcentaje de aciertos en el 20% superior seleccionando distintas variables mediante ε.

Las mejores veinte variables, de acuerdo con el criterio ε son:

N	*FIELD*
1	PPERSAUT
2	PBRAND
3	APLEZIER
4	MKOOPKLA
5	MOSTYPE
6	MOSHOOFD
7	APERSAUT
8	PWAPART
9	AWAPART
10	ABYSTAND
11	MHHUUR
12	MHKOOP
13	PPLEZIER
14	PMOTSCO
15	MOPLLAAG

N	*FIELD*
16	MOPLHOOG
17	MINKGEM
18	ABRAND
19	MSKA
20	MBERARBG

Tabla 3.24 Lista de 20 mejores variables usando el criterio ε.

Criterio ε'

El procedimiento es el mismo pero en esta ocasión se usan los valores de ε' para ordenar las variables. Como ε' sólo se obtiene para variables métricas, las variables MOSTYPE y MOSHOOFD no se toman en cuenta.

La siguiente gráfica muestra el desempeño usando distintos conjuntos de variables, desde 1 hasta 83 variables. El número de aciertos se está contando para el 20% superior.

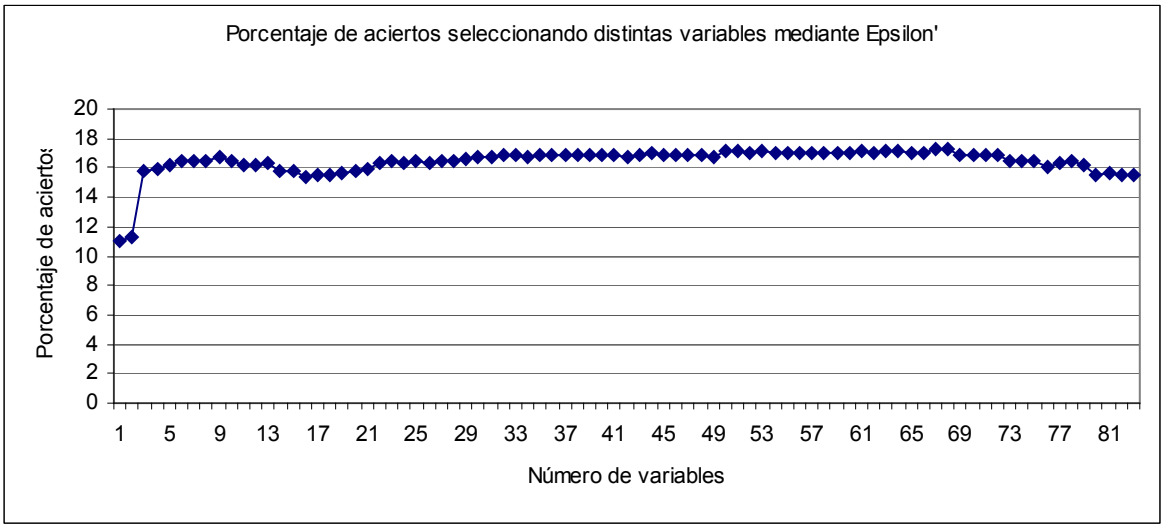

Fig. 3.8 Porcentaje de aciertos en el 20% superior seleccionando distintas variables mediante ε'.

Las variables 13 a 19 que bajan un poco el desempeño de la clasificación son variables sociodemográficas.

La variable con mejor valor de ε' es la misma variable con mejor valor de ε: PPERSAUT; por lo que el desempeño para una variable coincide con el de ε. Las tres variables con mayor valor de ε' (PPERSAUT, APERSAUT y PBRAND) proporcionaron un buen desempeño.

La siguiente tabla muestra los porcentajes de aciertos para los conjuntos de 1 a 20 variables.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
11.00	11.25	15.72	15.89	16.15	16.41	16.49	16.41	16.75	16.41	16.24	16.15	16.32	15.81	15.72	15.38	15.46	15.46	15.64	15.72

Tabla 3.25 Porcentaje de aciertos en el 20% superior seleccionando distintas variables mediante ε'.

Las mejores veinte variables, de acuerdo con el criterio ε' son:

N	*Variable*
1	PPERSAUT
2	APERSAUT
3	PBRAND
4	MKOOPKLA
5	PWAPART
6	MINKGEM
7	AWAPART
8	MRELGE
9	MHKOOP
10	MAUT1
11	MOPLHOOG
12	ABRAND
13	MINK4575
14	MSKA
15	MBERHOOG
16	MZPART
17	MINK7512
18	MBERMIDD
19	MOPLMIDD
20	APLEZIER

Tabla 3.26 Lista de 20 mejores variables usando el criterio ε'.

Criterio $p(C|\boldsymbol{X})$

La siguiente gráfica muestra el desempeño de la clasificación, sobre los datos de entrenamiento, utilizando distintos números de variables, desde una variable, hasta 85 variables. Para la clasificación usando solamente una variable se eligió la variable con el valor más alto de probabilidad $p(C|\boldsymbol{X})$ de todos sus posibles valores. Para la clasificación usando dos variables se eligieron las dos variables con los valores más altos de $p(C|\boldsymbol{X})$ de todos sus posibles valores y así sucesivamente hasta llegar a las 85 variables. El número de aciertos se está contando en el 20% superior.

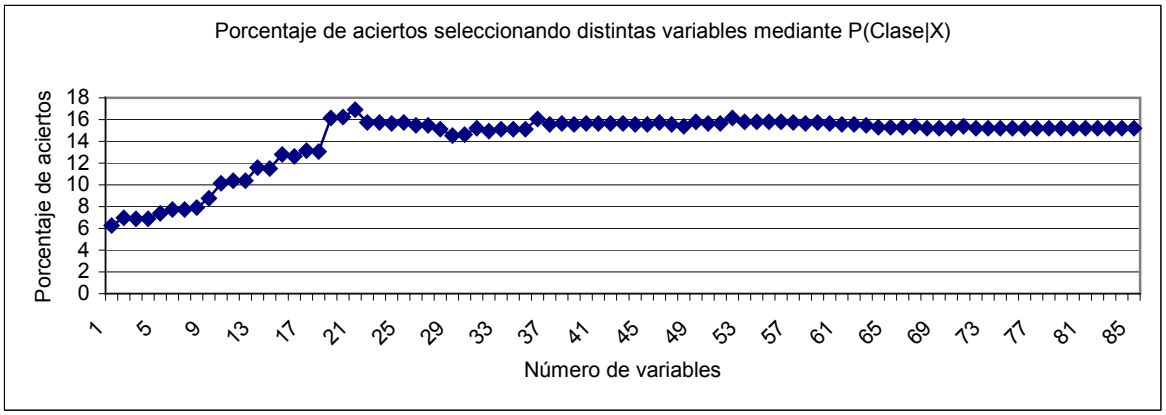

Fig. 3.9 Porcentaje de aciertos en el 20% superior seleccionando distintas variables mediante *p*(*C*|*X***).**

En la gráfica se observa que el desempeño usando solamente la primera variable es casi el mismo que si se seleccionaran las muestras (en este caso de estudio, los clientes) aleatoriamente y se alcanza el desempeño máximo usando las primeras 21 variables. A partir de ahi el desempeño ya no mejora. El mal desempeño para una variable se debe a que el primer indicador de *p*(*C*|***X***), *p*(*C*|PMOTSCO = 3), sólo se cumple para tres registros del conjunto de entrenamiento, y los demás indicadores de PMOTSCO tienen valores bajos de probabilidad P(*C*|***X***): *p*(*C*|PMOTSCO = 5) = 0.125, *p*(*C*|PMOTSCO = 4) = 0.066, *p*(*C*|PMOTSCO = 0) = 0.059, *p*(*C*|PMOTSCO = 6) = 0.020, *p*(*C*|PMOTSCO = 7) = 0 y para los demás valores de PMOTSCO (1, 2, 8, 9) no hay registros que tengan dichos valores. En la gráfica se observa como el desempeño va mejorando conforme se consideran más variables. Esto es debido a que los mejores indicadores de *p*(*C*|***X***) sólo se cumplen para un número limitado de registros por eso conforme se tienen más indicadores se van teniendo más registros con una buena probabilidad de pertenecer a la clase y se alcanza un límite cuando se consideran 21 variables.

La siguiente tabla muestra los porcentajes de aciertos en el 20% superior para los conjuntos de una a 21 variables.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*
6.27	6.96	6.87	6.87	7.39	7.73	7.73	7.90	8.76	10.14	10.40	10.40	11.60	11.51	12.80	12.63	13.14	13.06	16.15	16.24	16.92

Tabla 3.27 Porcentaje de aciertos en el 20% superior seleccionando distintas variables mediante *p*(*C*|*X***).**

Las mejores veintiún variables, de acuerdo con el criterio *p*(*C*|***X***) son:

N	*Variable*
1	PMOTSCO
2	PPLEZIER
3	PZEILPL
4	APLEZIER
5	AFIETS
6	ALEVEN
7	AZEILPL
8	PGEZONG
9	PBYSTAND

N	*Variable*
10	MINK7512
11	PWAOREG
12	AWAOREG
13	MSKB1
14	ABYSTAND
15	PWAPART
16	PINBOED
17	MSKA
18	AGEZONG
19	APERSAUT
20	MBERHOOG
21	MOSTYPE

Tabla 3.28 Lista de 21 mejores variables usando el criterio $p(C|X)$.

3.3.3 Clasificación del conjunto de entrenamiento usando diversos conjuntos de variables

En la sección anterior se usaron tres criterios para seleccionar las variables y analizar la clasificación hecha con los tres criterios. En esta sección se realizará la clasificación de los datos de entrenamiento mediante diversos conjuntos de variables con base en las características de las variables. Este enfoque multiperspectiva es útil para determinar si hay algún tipo de variables que sea más predictivo.

Las variables de los datos CoIL se pueden agrupar en tres distintos tipos: sociodemográficas, de contribuciones de pólizas y de números de pólizas. En secciones anteriores se estudiaron las características de las variables y se observó que las variables sociodemográficas tienen mayor varianza que las variables de contribuciones de pólizas y de números de pólizas. Debido a esta mayor varianza, uno esperaría que clasificando los datos usando sólo las variables sociomodegráficas se obtendrían mejores resultados que usando sólamente las variables de contribuciones o de números de pólizas. Sin embargo, al realizar estas clasificaciones no sucedió así. La razón de esto es que las variables sociodemográficas no son específicas a cada individuo, sino a las regiones geográficas donde viven los individuos, por lo que no proporcionan buen detalle a nivel individual.

En la siguiente tabla se muestran el porcentaje de aciertos sobre el 20% superior de los datos de entrenamiento. La clasificación se hizo utilizando los siguientes conjuntos de variables:

- Variables sociodemográficas
- Variables de contribuciones de pólizas
- Variables de números de pólizas
- Variables sociodemográficas y de contribuciones de pólizas
- Variables sociodemográficas y de números de pólizas
- Variables de contribuciones de pólizas y de números de pólizas

Contribuciones de pólizas	*Números de pólizas*	*Sociodemográficas*	*Contribuciones y números de pólizas*	*Sociodemográficas y contribuciones de pólizas*	*Sociodemográficas y números de pólizas*	*Todas las variables*
17.44	13.92	12.63	16.58	14.43	13.92	13.92

Tabla 3.29 Porcentajes de aciertos en el 20% superior usando diversos conjuntos de variables.

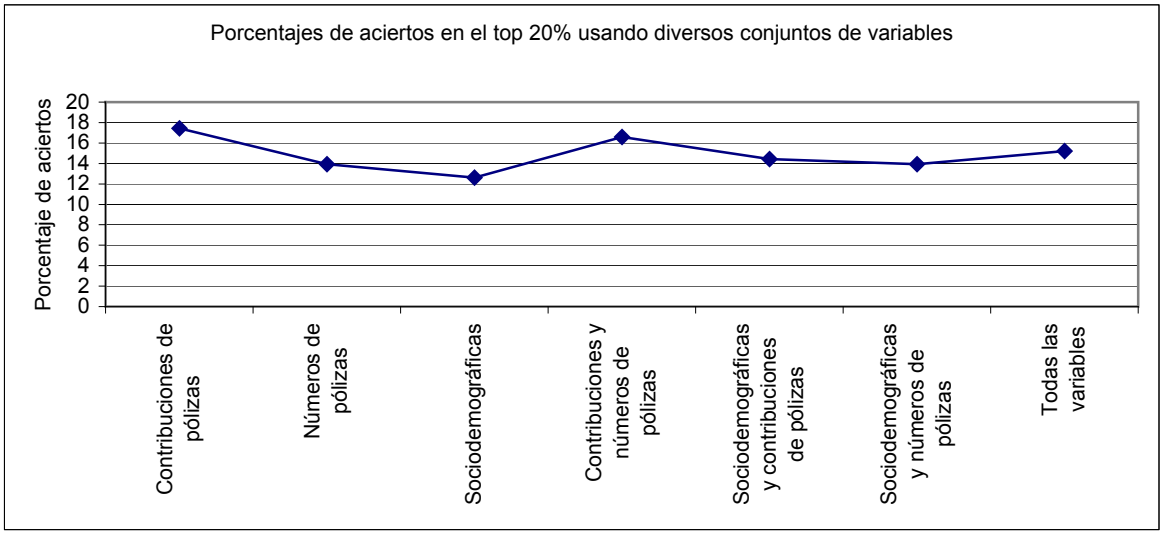

Fig. 3.10 Porcentaje de aciertos en el 20% superior usando diversos conjuntos de variables.

En la figura 3.10 se observa que las variables de contribuciones de pólizas proporcionaron un mejor desempeño que las variables de números de pólizas y sociodemográficas. También se puede apreciar que el desempeño de las variables de contribuciones junto con las variables de números de pólizas fue menor que usando solamente las variables de contribuciones. La razón de que el conjunto de variables de contribuciones de pólizas proporcione un buen desempeño puede ser que estas variables presentan una mayor varianza que las variables de números de pólizas y, a diferencia de las variables sociodemográficas, sus valores varían por individuo, no por grupos de individuos. También se observa que usando solamente las variables de contribuciones de pólizas se obtuvieron mejores resultados que usando todas las 85 variables.

Si se realiza nuevamente la clasificación Bayesiana ingenua, eligiendo las variables mediante los tres criterios, pero esta vez usando solamente las variables de contribuciones se obtienen los siguientes resultados:

Número de Variables	Variables ε	Desempeño ε	Variables ε'	Desempeño ε'	Variables $p(C\|\boldsymbol{X})$	Desempeño $p(C\|\boldsymbol{X})$
1	PPERSAUT	11.00	PPERSAUT	11.00	PMOTSCO	6.27
2	PBRAND	15.98	PBRAND	15.98	PPLEZIER	6.96
3	PWAPART	16.41	PWAPART	16.41	PZEILPL	6.87

Número de Variables	Variables ε	Desempeño ε	Variables ε'	Desempeño ε'	Variables $p(C\|\boldsymbol{X})$	Desempeño $p(C\|\boldsymbol{X})$
1	PPERSAUT	11.00	PPERSAUT	11.00	PMOTSCO	6.27
4	PPLEZIER	17.27	PPLEZIER	17.27	PGEZONG	7.22
5	PMOTSCO	17.18	PBYSTAND	17.27	PBYSTAND	7.73
6	PGEZONG	17.10	PFIETS	17.10	PWAOREG	7.90
7	PBYSTAND	17.10	PGEZONG	17.01	PWAPART	10.74
8	PWAOREG	17.10	PLEVEN	16.84	PINBOED	10.74
9	PZEILPL	17.01	PWAOREG	16.84	PAANHANG	10.65
10	PLEVEN	16.92	PAANHANG	16.84	PWABEDR	10.82
11	PFIETS	16.67	PZEILPL	16.84	PBRAND	13.83
12	PINBOED	16.75	PINBOED	16.84	PLEVEN	14.00
13	PAANHANG	16.75	PMOTSCO	16.75	PPERSAUT	16.84
14	PWABEDR	16.75	PWABEDR	16.75	PTRACTOR	17.10
15	PBROM	16.92	PBESAUT	16.75	PFIETS	16.92
16	PTRACTOR	17.18	PTRACTOR	16.92	PBROM	17.18
17	PWALAND	17.44	PPERSONG	16.92	PWALAND	17.44
18	PWERKT	17.44	PWALAND	17.18	PWERKT	17.44
19	PBESAUT	17.44	PVRAAUT	17.18	PBESAUT	17.44
20	PPERSONG	17.44	PWERKT	17.18	PPERSONG	17.44
21	PVRAAUT	17.44	PBROM	17.44	PVRAAUT	17.44

Tabla 3.30 Desempeño de diversos conjuntos de variables de contribuciones usando distintos criterios para formar los conjuntos de variables.

En los resultados de desempeño se observa que las tres variables que más contribuyen en un mejor desempeño son: PPERSAUT, PBRAND, PWAPART. Otras variables que también mejoraron el desempeño fueron: PPLEZIER, PBROM, PTRACTOR y PWALAND.

Después de observar estos resultados parece natural utilizar las variables PPERSAUT, PBRAND, PWAPART, PPLEZIER, PBROM, PTRACTOR y PWALAND para clasificar el conjunto de entrenamiento e ir agregando variables sociodemográficas a estas variables para analizar si se mejora el desempeño.

Usando las siete variables mencionadas se obtuvo un desempeño de 17.53% en el 20% superior de los datos de entrenamiento, mayor a los desempeños obtenidos anteriormente.

Una manera para determinar cuáles son las variables sociodemográficas que conviene elegir es analizando el desempeño que tiene cada variable de manera individual en la elección del 20% superior de los datos de entrenamiento.

La siguiente tabla muestra los mejores diez desempeños de las variables sociodemográficas en el 20% superior del conjunto de entrenamiento.

N	Variable	Desempeño
1	MOSTYPE	11.51
2	MOSHOOFD	10.40
3	MKOOPKLA	10.40
4	MOPLLAAG	9.79

N	Variable	Desempeño
5	MINKGEM	9.79
6	MHHUUR	9.71
7	MHKOOP	9.71
8	MSKA	9.54
9	MOPLHOOG	9.45
10	MSKC	9.28

Tabla 3.31 Mejores desempeños de variables sociomodegráficas.

De acuerdo con la lista las variables MOSTYPE, MOSHOOFD y MKOOPKLA son buenas candidatas para agregarlas al conjunto inicial de seis variables.

Se hicieron pruebas agregando estas tres variables por separado. Agregando MOSTYPE se obtuvo un desempeño de 17.01% en el 20% superior, agregando MOSHOOFD se obtuvo un desempeño de 16.58% en el 20% superior y agregando MKOOPKLA se obtuvo un desempeño de 17.18%. Agregando las tres variables se obtuvo un desempeño de 16.49%. El desempeño sin las variables sociodemográficas es del 17.53%, así que en los tres casos, el desempeño disminuyó un poco.

En su artículo *"Magical Thinking in Data Mining"* [9], Elkan menciona que la exactitud predictiva de la clasificación Bayesiana ingenua generalmente se puede mejorar mediante técnicas de *boosting*, así como agregando nuevas variables derivadas de las originales ya que son maneras de relajar el supuesto de que las variables son independientes. Elkan realizó pruebas añadiendo dos variables derivadas, las cuales mejoraron el desempeño de la clasificación Bayesian ingenua. También realizó pruebas de *boosting* con estas dos nuevas variables pero el desempeño no mejoró significativamente, razón por la cual lo descartó.

Las dos variables derivadas que se agregaron proporcionan información más detallada sobre las pólizas de auto y contra incendios de un individuo. Estas variables derivadas consistieron en el producto cruz de las dos variables correspondientes a pólizas de auto (PPERSAUT y APERSAUT) y de las dos variables correspondientes a las pólizas contra incendios (PBRAND y ABRAND). Llamaremos a estas dos variables PERSAUT_CROSS y BRAND_CROSS.

La siguiente gráfica muestra como mejoraron los desempeños empleando estas dos variables en los mismos conjuntos de variables que se usaron anteriormente.

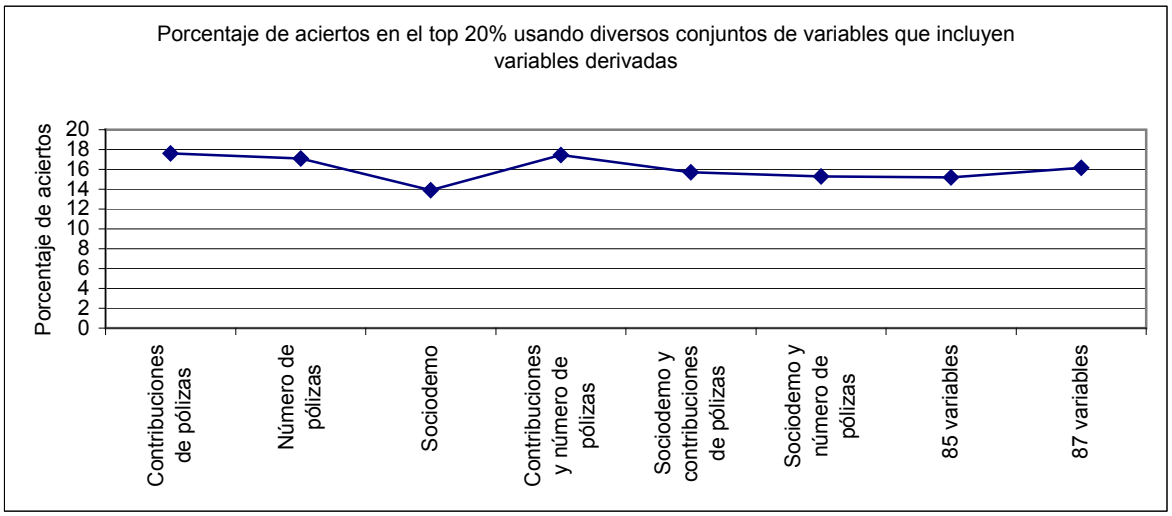

Fig. 3.11 Porcentaje de aciertos en el 20% superior usando diversos conjuntos de variables que incluyen las dos variables derivadas.

Los resultados mejoraron en todos los conjuntos de variables usados. La razón de la mejora se debe a que el crear variables derivadas que son producto cruz de otras variables permite a la clasificación Bayesiana ingenua asociar una probabilidad diferente de póliza de casa rodante con cada par de valores de PPERSAUT y APERSAUT, así como con cada par de valores de PBRAND y ABRAND. El motivo de por qué se obtienen variables derivadas únicamente a partir de los atributos mencionados es porque estas variables son las que presentan mayor varianza y por lo mismo tiene sentido desmenuzarlas en mayor detalle.

La siguiente tabla compara los desempeños usando los diversos conjuntos de variables.

Conjuntos de variables	Desempeño en el 20% superior sin variables derivadas	Desempeño 20% superior con variables derivadas'
Contribuciones de pólizas	17.44	17.61
Números de pólizas	13.92	17.10
Sociodemográficas	12.63	13.92
Contribuciones y números de pólizas	16.58	17.44
Sociodemográficas y contribuciones de pólizas	14.43	15.72
Sociodemográficas y números de pólizas	13.92	15.29
PPERSAUT,PBRAND,PWAPART,PPLEZIER, PBROM, PTRACTOR y PWALAND	17.53	17.35
Todas las variables	15.21	16.15

Tabla 3.32 Desempeño de diversos conjuntos de variables de contribuciones, con y sin variables derivadas.

Las variables PPERSAUT, PBRAND, PWAPART, PPLEZIER, PBROM, PTRACTOR y PWALAND mostradas en la tabla 3.32 fueron las mejores variables de contribuciones elegidas de la tabla 3.30.

En la tabla se observa que los mejores resultados usando variables derivadas fueron para los conjuntos:

- Contribuciones de pólizas
- Contribuciones y números de pólizas
- Siete mejores variables (PPERSAUT, PBRAND, PWAPART, PPLEZIER, PBROM, PTRACTOR y PWALAND)

Sobre estos tres conjuntos se agregaron las variables sociodemográficas y se obtuvieron los siguientes resultados:

Conjuntos de variables	Desempeño 20% superior
Contribuciones de pólizas + derivadas + MOSTYPE	17.70
Contribuciones de pólizas + derivadas + MOSHOOFD	17.61
Contribuciones de pólizas + derivadas + MKOOPKLA	17.87
Contribuciones y números de pólizas + derivadas + MOSTYPE	17.61
Contribuciones y números de pólizas + derivadas + MOSHOOFD	17.61
Contribuciones y números de pólizas + derivadas + MKOOPKLA	17.53
Siete mejores variables + derivadas + MOSTYPE	17.70
Siete mejores variables + derivadas + MOSHOOFD	17.35
Siete mejores variables + derivadas + MKOOPKLA	17.87

Tabla 3.33 Desempeño de diversos conjuntos de variables de contribuciones, añadiendo variables sociodemográficas.

A diferencia de cuando no se incluyen las variables derivadas, los desempeños al agregar por separado las tres variables sociodemográficas sí mejoraron. De estas tres variables sociodemográficas las que tuvieron mejor influencia en el desempeño fueron las variables MOSTYPE y MKOOPLA.

Con base en los análisis previos se determina que los mejores conjuntos de variables para obtener buen desempeño en el 20% superior de los datos de entrenamiento son:

- 21 variables de contribuciones + 2 variables derivadas + MOSTYPE: 17.70%
- 21 variables de contribuciones + 2 variables derivadas + MKOOPKLA: 17.87%
- 7 mejores variables + 2 variables derivadas + MOSTYPE: 17.70%
- 7 mejores variables + 2 variables derivadas +MKOOPKLA. 17.87%

3.3.4 Clasificación del conjunto de entrenamiento mediante la función ε y las probabilidades $p(\boldsymbol{C}|X)$

En la sección de ε y ε' se mencionó cómo se puede usar la función ε para clasificar los datos. El mismo procedimiento se puede emplear usando la probabilidad $p(C|\boldsymbol{X})$ en lugar de la función ε. La función ε' no se puede emplear porque ε' corresponde a una variable, no a una variable y sus valores, por lo que no hay manera de asociar un conjunto de individuos a la función ε'.

Clasificación mediante la función ε

Se seleccionó el 20% superior de los clientes con mayor probabilidad de comprar una póliza de seguro de casa rodante mediante la función ε. De esta manera se obtuvo un porcentaje de aciertos de 11% en el 20% superior de los datos de entrenamiento. La razón de que no sea un buen porcentaje se debe a que el mejor valor de ε, que corresponde a PPERSAUT = 6, se cumple para 2319 registros (39% de los datos de entrenamiento) y la probabilidad $p(C|\text{PPERSAUT} = 6)$ es 11.30%, así que usando este indicador era de esperarse que el porcentaje de aciertos del 20% superior no fuera mayor del 11.30%.

Una forma de mejorar este desempeño es mediante un enfoque multiperspectiva. En el caso de los indicadores de ε este enfoque se puede aplicar tomando en cuenta el número de indicadores que dispara el individuo. De esta manera, se le asignará mayor peso a un individuo que tiene un alto valor de ε y que dispara más de un indicador que a un individuo con el mismo valor de ε pero que dispara menos indicadores. Aplicando este enfoque se obtuvo un porcentaje de aciertos de 16.24% en el 20% superior, lo cual fue una mejora significativa.

Clasificación mediante la probabilidad $p(C|\boldsymbol{X})$

Usando la probabilidad $p(C|\boldsymbol{X})$, en lugar de la función ε, se obtuvo un porcentaje de aciertos de 13.23% en el 20% superior de los datos de entrenamiento. Cuando se aplicó el enfoque multiperspectiva el desempeñó no mejoró. Esto se debe a que el enfoque multiperspectiva empleado no cambió la forma en que se eligieron los individuos del 20% superior ya que los mejores indicadores de $p(C|\boldsymbol{X})$ se cumplen para pocos individuos. El enfoque multiperspectiva empleado mejora los resultados en los casos en que los indicadores se cumplen para un número grande de individuos.

Aunque los resultados de la clasificación mediante los indicadores fueron buenos y mejoraron al emplear un enfoque multiperspectiva, los desempeños no fueron mejores que los obtenidos mediante la clasificación Bayesiana ingenua, por ello, sólo se empleará el modelo de clasificación Bayesiana ingenua para clasificar los datos de prueba.

3.3.5 Clasificación del conjunto de prueba mediante clasificación Bayesiana ingenua

Los resultados de la clasificación en el conjunto de prueba fueron los siguientes:

Conjuntos de variables	Desempeño 20% superior
21 variables de contribuciones + 2 variables derivadas + MOSTYPE	15.00%
21 variables de contribuciones + 2 variables derivadas + MKOOPKLA	15.13%
7 mejores variables + 2 variables derivadas + MOSTYPE	14.38%
7 mejores variables + 2 variables derivadas + MKOOPKLA	14.13%

Tabla 3.34 Porcentajes de aciertos en el 20% superior de los datos de prueba.

De estos resultados el mejor fue el obtenido con las 21 variables de contribuciones junto con las dos variables derivadas y la variable sociodemográfica MKOOPKLA. Este resultado es igual al del ganador del *Challenge* 2000 (121 aciertos en los 800 individuos elegidos). Las variables usadas por Elkan fueron las 42 variables de contribuciones y

números de pólizas junto con las dos variables derivadas y la variable sociodemográfica MKOOPKLA. El añadir las dos variables derivadas proporcionó un mejor desempeño.

También es de interés hacer notar que en el concurso *Challenge* 2000 los dos primeros lugares de predicción usaron la clasificación Bayesiana ingenua [9]. La razón de que la clasificación Bayesiana ingenua obtenga buenos resultados se puede deber a que aminora los problemas de tener poca certeza estadística al involucrar más muestras en el cálculo de las probabilidades $p(C|\boldsymbol{X})$.

Durante este capítulo se mostró como el uso de un enfoque multiperspectiva permite resolver diferentes objetivos del problema. Este fue el caso al analizar por separado los distintos tipos de variables: el uso de las variables de contribuciones de pólizas es conveniente para realizar predicciones, mientras que el uso de variables sociodemográficas es útil para identificar zonas geográficas con clientes potenciales. El enfoque multiperspectiva también se aplicó al combinar dos variables distintas mediante el producto cruz en una nueva variable derivada. Al incorporar en el modelo dos nuevas variables obtenidas de esta manera se mejoró el desempeño de la clasificación.

Capítulo 4: Caso de estudio "Predicción de pacientes más costosos"

Los altos costos en el sector salud repercuten de manera negativa en la prosperidad económica de un país. Debido a esto es importante contar con modelos de predicciones de costos que puedan utilizarse para administrar mejor los recursos y las medidas preventivas de salud. Una encuesta de 1996, realizada por el gobierno de Estados Unidos, reveló que el 1% más costoso de la población usaba el 27% de los recursos, y el 5% más costoso usaba el 55% [4]. La distribución sesgada de los costos de salud muestra la importancia de identificar a los individuos más probables de incurrir en gastos de salud que sean muy altos.

Debido a esta necesidad, existen compañías como DxCG (http://www.dxcg.com), la cual es una compañía de software de modelos de predicción. Uno de los modelos de DxCG tiene como propósito predecir los costos de un individuo para el año siguiente con base en datos del individuo del año actual. Adaptive Technologies (http://www.at-inc.biz) realizó una investigación para mejorar el modelo de DxCG usando técnicas basadas en algoritmos evolutivos, pero sin comprometer la transparencia y comprensión del enfoque actual usado por DxCG, el cual es un modelo de regresión lineal que toma en cuenta variables asociadas con la condición médica del paciente así como algunas variables sociodemográficas.

El objetivo específico del modelo a desarrollar fue predecir el 0.5% de los pacientes más costosos en el siguiente año dados los datos de los individuos en el año actual. Para medir el desempeño de las predicciones se obtuvo el porcentaje de aciertos en el 0.5% de individuos predicho.

4.1 Descripción de los datos

Los datos proporcionados por DxCG fueron:

1. Variables sociodemográficas, de costos médicos y de condición médica de pacientes en el año 1997 y los costos que tuvieron en el año 1998. Este conjunto tiene 29,062 registros de pacientes.
2. Variables sociodemográficas, de costos médicos y de condición médica de pacientes en el año 1998 y los costos que tuvieron en el año 1999. Este conjunto tiene 38,879 registros de pacientes.
3. Variables sociodemográficas, de costos médicos y de condición médica de los individuos en el año 2000. Este conjunto tiene 90,104 registros de pacientes.

Los dos primeros conjuntos de datos se usaron para entrenar el modelo y la predicción se realizó sobre el conjunto de pacientes del año 2000. Después de entregar los resultados de la predicción de los pacientes del 2000, DxCG proporcionó los costos del 2001 de los mismos individuos para medir el desempeño de la predicción.

Los datos proporcionados por DxCG fueron obtenidos del MEDSTAT (Marketscan Research Database) para un grupo de personas diagnosticadas con diabetes. Cada conjunto de datos consiste de variables asociadas a costos de los pacientes, así como a su condición médica y otras variables como edad, sexo, número de hospitalizaciones y

número de días pasados en el hospital. A partir de las variables proporcionadas por DxCG se crearon otras variables, las cuales se mencionarán más adelante.

Es importante mencionar que los grupos de pacientes son diferentes para los tres años dados.

Los costos fueron dados para cada trimestre del año y por tipo de costo (gastos internos, gastos externos y gastos de medicinas). Las condiciones médicas consisten en 184 variables denominadas HCCs (Hierarchical Condition Categories).

Para reducir la dimensión de los datos, se realizó *coarse graining* de las variables que tenían un número grande de valores, por ejemplo la edad, y de las variables que no eran discretas, principalmente las variables asociadas a costos.

4.1.1 Variables de los datos

La siguiente tabla muestra las variables de los datos DxCG

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
1	YEAR	*Concurrent year*	No	0	Numérico
2	ENROLID	*Enrollee ID*	No	0	Numérico
3	SEX	*Sex*	Sí	2	T1
4	AGE	*Person age in years*	Sí	8	T2
5	ELIG1	*Number of months eligible in year 1*	Sí	4	T3
6	ELIG2	*Number of months eligible in year 2*	Sí	4	T3
7	ACOVY	*Total spending I+O+D*	Sí	6	T4
8	COVYI	*Total inpatient spending*	Sí	6	T4
9	COVYO	*Total outpatient spending*	Sí	6	T4
10	COVYD	*Total drug spending*	Sí	6	T4
11	X_1*	*Q1 costs normalized to the DXCG estimate*	Sí	6	T4
12	X_2*	*Q2 costs normalized to the DXCG estimate*	Sí	6	T4
13	X_3*	*Q3 costs normalized to the DXCG estimate*	Sí	6	T4
14	X_4*	*Q4 costs normalized to the DXCG estimate*	Sí	6	T4
15	SLOPE*	*Linear regression slope*	Sí	3	T5
16	OFFSET*	*Linear regression offset*	Sí	2	T6
17	SUDDENNESS*	*Surge in costs in the fourth quarter*	Sí	3	T7
18	CONCAVITY*	*Concavity of the quarterly costs*	Sí	3	T7
19	RMSE*	*Root mean squared error of the linear fit to quarterly costs*	Sí	5	T8
20	NLR*	*Non-linearity of the quarterly costs*	Sí	5	T8
21	COVYI_1	*Q1 total inpatient spending*	Sí	6	T4

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
22	COVYI_2	*Q2 total inpatient spending*	Sí	6	T4
23	COVYI_3	*Q3 total inpatient spending*	Sí	6	T4
24	COVYI_4	*Q4 total inpatient spending*	Sí	6	T4
25	TLOS	*Total length of stay*	Sí	8	T9
26	NHOSP	*Number of hospitalization*	Sí	4	T10
27	HOS1*	*Number of hospitalization days in Q1*	Sí	6	T11
28	HOS2*	*Number of hospitalization days in Q2*	Sí	6	T11
29	HOS3*	*Number of hospitalization days in Q3*	Sí	6	T11
30	HOS4*	*Number of hospitalization days in Q4*	Sí	6	T11
31	COVYO_1	*Q1 total outpatient spending*	Sí	6	T4
32	COVYO_2	*Q2 total outpatient spending*	Sí	6	T4
33	COVYO_3	*Q3 total outpatient spending*	Sí	6	T4
34	COVYO_4	*Q4 total outpatient spending*	Sí	6	T4
35	COVYD_1	*Q1 total drug spending*	Sí	6	T4
36	COVYD_2	*Q2 total drug spending*	Sí	6	T4
37	COVYD_3	*Q3 total drug spending*	Sí	6	T4
38	COVYD_4	*Q4 total drug spending*	Sí	6	T4
39	COSTRI4*	*Change in total spending from Q1, Q2, Q3 to Q4*	Sí	6	T12
40	COSTRI43*	*Change in total spending from Q1, Q2 to Q3, Q4*	Sí	6	T12
41	COSTRII4*	*Change in total inpatient spending from Q1, Q2, Q3 to Q4*	Sí	6	T12
42	COSTRII43*	*Change in total inpatient spending from Q1, Q2 to Q3, Q4*	Sí	6	T12
43	ACOVY_1	*Q1 total spending I+O+D*	Sí	6	T4
44	ACOVY_2	*Q2 total spending I+O+D*	Sí	6	T4
45	ACOVY_3	*Q3 total spending I+O+D*	Sí	6	T4
46	ACOVY_4	*Q4 total spending I+O+D*	Sí	6	T4
47	COSTDXCG	*Cost estimate by DXCG model*	Sí	6	T4
48	SCORE*	*Score function*	Sí	6	T4
49	HCC001	*HIV/AIDS*	Sí	2	Binario
50	HCC002	*Septicemia/Shock*	Sí	2	Binario
51	HCC003	*Central Nervous System Infection*	Sí	2	Binario
52	HCC004	*Tuberculosis*	Sí	2	Binario
53	HCC005	*Opportunistic Infections*	Sí	2	Binario
54	HCC006	*Other Infectious Diseases*	Sí	2	Binario
55	HCC007	*Metastatic Cancer and Acute Leukemia*	Sí	2	Binario

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
56	HCC008	*Lung, Upper Digestive Tract, and Other Severe Cancers*	Sí	2	Binario
57	HCC009	*Lymphatic, Head and Neck, Brain, and Other Major Cancers*	Sí	2	Binario
58	HCC010	*Breast, Prostate, Colorectal and Other Cancers and Tumors*	Sí	2	Binario
59	HCC011	*Other Respiratory and Heart Neoplasms*	Sí	2	Binario
60	HCC012	*Other Digestive and Urinary Neoplasms*	Sí	2	Binario
61	HCC013	*Other Neoplasms*	Sí	2	Binario
62	HCC014	*Benign Neoplasms of Skin, Breast, Eye*	Sí	2	Binario
63	HCC015	*Diabetes with Renal Manifestation*	Sí	2	Binario
64	HCC016	*Diabetes with Neurologic or Peripheral Circulatory Manifestation*	Sí	2	Binario
65	HCC017	*Diabetes with Acute Complications*	Sí	2	Binario
66	HCC018	*Diabetes with Ophthalmologic Manifestation*	Sí	2	Binario
67	HCC019	*Diabetes with No or Unspecified Complications*	Sí	2	Binario
68	HCC020	*Type I Diabetes Mellitus*	Sí	2	Binario
69	HCC021	*Protein-Calorie Malnutrition*	Sí	2	Binario
70	HCC022	*Other Significant Endocrine and Metabolic Disorders*	Sí	2	Binario
71	HCC023	*Disorders of Fluid/Electrolyte/Acid-Base Balance*	Sí	2	Binario
72	HCC024	*Other Endocrine/Metabolic/Nutritional Disorders*	Sí	2	Binario
73	HCC025	*End-Stage Liver Disease*	Sí	2	Binario
74	HCC026	*Cirrhosis of Liver*	Sí	2	Binario
75	HCC027	*Chronic Hepatitis*	Sí	2	Binario
76	HCC028	*Acute Liver Failure/Disease*	Sí	2	Binario
77	HCC029	*Other Hepatitis and Liver Disease*	Sí	2	Binario
78	HCC030	*Gallbladder and Biliary Tract Disorders*	Sí	2	Binario
79	HCC031	*Intestinal Obstruction/Perforation*	Sí	2	Binario
80	HCC032	*Pancreatic Disease*	Sí	2	Binario
81	HCC033	*Inflammatory Bowel Disease*	Sí	2	Binario
82	HCC034	*Peptic Ulcer, Hemorrhage, Other Specified*	Sí	2	Binario

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
		Gastrointestinal Disorders			
83	HCC035	*Appendicitis*	Sí	2	Binario
84	HCC036	*Other Gastrointestinal Disorders*	Sí	2	Binario
85	HCC037	*Bone/Joint/Muscle Infections/Necrosis*	Sí	2	Binario
86	HCC038	*Rheumatoid Arthritis and Inflammatory Connective Tissue Disease*	Sí	2	Binario
87	HCC039	*Disorders of the Vertebrae and Spinal Discs*	Sí	2	Binario
88	HCC040	*Osteoarthritis of Hip or Knee*	Sí	2	Binario
89	HCC041	*Osteoporosis and Other Bone/Cartilage Disorders*	Sí	2	Binario
90	HCC042	*Congenital/Developmental Skeletal and Connective Tissue Disorders*	Sí	2	Binario
91	HCC043	*Other Musculoskeletal and Connective Tissue Disorders*	Sí	2	Binario
92	HCC044	*Severe Hematological Disorders*	Sí	2	Binario
93	HCC045	*Disorders of Immunity*	Sí	2	Binario
94	HCC046	*Coagulation Defects and Other Specified Hematological Disorders*	Sí	2	Binario
95	HCC047	*Iron Deficiency and Other/Unspecified Anemias and Blood Disease*	Sí	2	Binario
96	HCC048	*Delirium and Encephalopathy*	Sí	2	Binario
97	HCC049	*Dementia*	Sí	2	Binario
98	HCC050	*Senility, Nonpsychotic Organic Brain Syndromes/Conditions*	Sí	2	Binario
99	HCC051	*Drug/Alcohol Psychosis*	Sí	2	Binario
100	HCC052	*Drug/Alcohol Dependence*	Sí	2	Binario
101	HCC053	*Drug/Alcohol Abuse, Without Dependence*	Sí	2	Binario
102	HCC054	*Schizophrenia*	Sí	2	Binario
103	HCC055	*Major Depressive, Bipolar, and Paranoid Disorders*	Sí	2	Binario
104	HCC056	*Reactive and Unspecified Psychosis*	Sí	2	Binario
105	HCC057	*Personality Disorders*	Sí	2	Binario
106	HCC058	*Depression*	Sí	2	Binario
107	HCC059	*Anxiety Disorders*	Sí	2	Binario
108	HCC060	*Other Psychiatric Disorders*	Sí	2	Binario
109	HCC061	*Profound Mental Retardation/Developmental Disability*	Sí	2	Binario

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
110	HCC062	*Severe Mental Retardation/Developmental Disability*	Sí	2	Binario
111	HCC063	*Moderate Mental Retardation/Developmental Disability*	Sí	2	Binario
112	HCC064	*Mild/Unspecified Mental Retardation/Developmental Disability*	Sí	2	Binario
113	HCC065	*Other Developmental Disability*	Sí	2	Binario
114	HCC066	*Attention Deficit Disorder*	Sí	2	Binario
115	HCC067	*Quadriplegia, Other Extensive Paralysis*	Sí	2	Binario
116	HCC068	*Paraplegia*	Sí	2	Binario
117	HCC069	*Spinal Cord Disorders/Injuries*	Sí	2	Binario
118	HCC070	*Muscular Dystrophy*	Sí	2	Binario
119	HCC071	*Polyneuropathy*	Sí	2	Binario
120	HCC072	*Multiple Sclerosis*	Sí	2	Binario
121	HCC073	*Parkinson's and Huntington's Diseases*	Sí	2	Binario
122	HCC074	*Seizure Disorders and Convulsions*	Sí	2	Binario
123	HCC075	*Coma, Brain Compression/Anoxic Damage*	Sí	2	Binario
124	HCC076	*Mononeuropathy, Other Neurological Conditions/Injuries*	Sí	2	Binario
125	HCC077	*Respirator Dependence/Tracheostomy Status*	Sí	2	Binario
126	HCC078	*Respiratory Arrest*	Sí	2	Binario
127	HCC079	*Cardio-Respiratory Failure and Shock*	Sí	2	Binario
128	HCC080	*Congestive Heart Failure*	Sí	2	Binario
129	HCC081	*Acute Myocardial Infarction*	Sí	2	Binario
130	HCC082	*Unstable Angina and Other Acute Ischemic Heart Disease*	Sí	2	Binario
131	HCC083	*Angina Pectoris/Old Myocardial Infarction*	Sí	2	Binario
132	HCC084	*Coronary Atherosclerosis/Other Chronic Ischemic Heart Disease*	Sí	2	Binario
133	HCC085	*Heart Infection/Inflammation, Except Rheumatic*	Sí	2	Binario
134	HCC086	*Valvular and Rheumatic Heart Disease*	Sí	2	Binario
135	HCC087	*Major Congenital*	Sí	2	Binario

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
		Cardiac/Circulatory Defect			
136	HCC088	*Other Congenital Heart/Circulatory Disease*	Sí	2	Binario
137	HCC089	*Hypertensive Heart and Renal Disease or Encephalopathy*	Sí	2	Binario
138	HCC090	*Hypertensive Heart Disease*	Sí	2	Binario
139	HCC091	*Hypertension*	Sí	2	Binario
140	HCC092	*Specified Heart Arrhythmias*	Sí	2	Binario
141	HCC093	*Other Heart Rhythm and Conduction Disorders*	Sí	2	Binario
142	HCC094	*Other and Unspecified Heart Disease*	Sí	2	Binario
143	HCC095	*Cerebral Hemorrhage*	Sí	2	Binario
144	HCC096	*Ischemic or Unspecified Stroke*	Sí	2	Binario
145	HCC097	*Precerebral Arterial Occlusion and Transient Cerebral Ischemia*	Sí	2	Binario
146	HCC098	*Cerebral Atherosclerosis and Aneurysm*	Sí	2	Binario
147	HCC099	*Cerebrovascular Disease, Unspecified*	Sí	2	Binario
148	HCC100	*Hemiplegia/Hemiparesis*	Sí	2	Binario
149	HCC101	*Diplegia (Upper), Monoplegia, and Other Paralytic Syndromes*	Sí	2	Binario
150	HCC102	*Speech, Language, Cognitive, Perceptual Deficits*	Sí	2	Binario
151	HCC103	*Cerebrovascular Disease Late Effects, Unspecified*	Sí	2	Binario
152	HCC104	*Vascular Disease with Complications*	Sí	2	Binario
153	HCC105	*Vascular Disease*	Sí	2	Binario
154	HCC106	*Other Circulatory Disease*	Sí	2	Binario
155	HCC107	*Cystic Fibrosis*	Sí	2	Binario
156	HCC108	*Chronic Obstructive Pulmonary Disease*	Sí	2	Binario
157	HCC109	*Fibrosis of Lung and Other Chronic Lung Disorders*	Sí	2	Binario
158	HCC110	*Asthma*	Sí	2	Binario
159	HCC111	*Aspiration and Specified Bacterial Pneumonias*	Sí	2	Binario
160	HCC112	*Pneumococcal Pneumonia, Empyema, Lung Abscess*	Sí	2	Binario
161	HCC113	*Viral and Unspecified Pneumonia, Pleurisy*	Sí	2	Binario
162	HCC114	*Pleural Effusion/Pneumothorax*	Sí	2	Binario
163	HCC115	*Other Lung Disorders*	Sí	2	Binario

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
164	HCC116	*Legally Blind*	Sí	2	Binario
165	HCC117	*Major Eye Infections/Inflammations*	Sí	2	Binario
166	HCC118	*Retinal Detachment*	Sí	2	Binario
167	HCC119	*Proliferative Diabetic Retinopathy and Vitreous Hemorrhage*	Sí	2	Binario
168	HCC120	*Diabetic and Other Vascular Retinopathies*	Sí	2	Binario
169	HCC121	*Retinal Disorders, Except Detachment and Vascular Retinopathies*	Sí	2	Binario
170	HCC122	*Glaucoma*	Sí	2	Binario
171	HCC123	*Cataract*	Sí	2	Binario
172	HCC124	*Other Eye Disorders*	Sí	2	Binario
173	HCC125	*Significant Ear, Nose, and Throat Disorders*	Sí	2	Binario
174	HCC126	*Hearing Loss*	Sí	2	Binario
175	HCC127	*Other Ear, Nose, Throat, and Mouth Disorders*	Sí	2	Binario
176	HCC128	*Kidney Transplant Status*	Sí	2	Binario
177	HCC129	*End Stage Renal Disease*	Sí	2	Binario
178	HCC130	*Dialysis Status*	Sí	2	Binario
179	HCC131	*Renal Failure*	Sí	2	Binario
180	HCC132	*Nephritis*	Sí	2	Binario
181	HCC133	*Urinary Obstruction and Retention*	Sí	2	Binario
182	HCC134	*Incontinence*	Sí	2	Binario
183	HCC135	*Urinary Tract Infection*	Sí	2	Binario
184	HCC136	*Other Urinary Tract Disorders*	Sí	2	Binario
185	HCC137	*Female Infertility*	Sí	2	Binario
186	HCC138	*Pelvic Inflammatory Disease and Other Specified Female Genital Disorders*	Sí	2	Binario
187	HCC139	*Other Female Genital Disorders*	Sí	2	Binario
188	HCC140	*Male Genital Disorders*	Sí	2	Binario
189	HCC141	*Ectopic Pregnancy*	Sí	2	Binario
190	HCC142	*Miscarriage/Abortion*	Sí	2	Binario
191	HCC143	*Completed Pregnancy With Major Complications*	Sí	2	Binario
192	HCC144	*Completed Pregnancy With Complications*	Sí	2	Binario
193	HCC145	*Completed Pregnancy Without Complications (Normal Delivery)*	Sí	2	Binario
194	HCC146	*Uncompleted Pregnancy With Complications*	Sí	2	Binario

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
195	HCC147	*Uncompleted Pregnancy With No or Minor Complications*	Sí	2	Binario
196	HCC148	*Decubitus Ulcer of Skin*	Sí	2	Binario
197	HCC149	*Chronic Ulcer of Skin, Except Decubitus*	Sí	2	Binario
198	HCC150	*Extensive Third-Degree Burns*	Sí	2	Binario
199	HCC151	*Other Third-Degree and Extensive Burns*	Sí	2	Binario
200	HCC152	*Cellulitis, Local Skin Infection*	Sí	2	Binario
201	HCC153	*Other Dermatological Disorders*	Sí	2	Binario
202	HCC154	*Severe Head Injury*	Sí	2	Binario
203	HCC155	*Major Head Injury*	Sí	2	Binario
204	HCC156	*Concussion or Unspecified Head Injury*	Sí	2	Binario
205	HCC157	*Vertebral Fractures*	Sí	2	Binario
206	HCC158	*Hip Fracture/Dislocation*	Sí	2	Binario
207	HCC159	*Major Fracture, Except of Skull, Vertebrae, or Hip*	Sí	2	Binario
208	HCC160	*Internal Injuries*	Sí	2	Binario
209	HCC161	*Traumatic Amputation*	Sí	2	Binario
210	HCC162	*Other Injuries*	Sí	2	Binario
211	HCC163	*Poisonings and Allegic Reactions*	Sí	2	Binario
212	HCC164	*Major Complications of Medical Care and Trauma*	Sí	2	Binario
213	HCC165	*Other Complications of Medical Care*	Sí	2	Binario
214	HCC166	*Major Symptoms, Abnormalities*	Sí	2	Binario
215	HCC167	*Minor Symptoms, Signs, Findings*	Sí	2	Binario
216	HCC168	*Extremely Low Birthweight Neonates*	Sí	2	Binario
217	HCC169	*Very Low Birthweight Neonates*	Sí	2	Binario
218	HCC170	*Serious Perinatal Problem Affecting Newborn*	Sí	2	Binario
219	HCC171	*Other Perinatal Problems Affecting Newborn*	Sí	2	Binario
220	HCC172	*Normal, Single Birth*	Sí	2	Binario
221	HCC173	*Major Organ Transplant*	Sí	2	Binario
222	HCC174	*Major Organ Transplant Status*	Sí	2	Binario
223	HCC175	*Other Organ Transplant/Replacement*	Sí	2	Binario
224	HCC176	*Artificial Openings for Feeding or Elimination*	Sí	2	Binario

Num Variable	Variable	Descripción	Métrica	Cardinalidad	Tipo Valor
225	HCC177	*Amputation Status, Lower Limb/Amputation Complications*	Sí	2	Binario
226	HCC178	*Amputation Status, Upper Limb*	Sí	2	Binario
227	HCC179	*Post-Surgical States/Aftercare/Elective*	Sí	2	Binario
228	HCC180	*Radiation Therapy*	Sí	2	Binario
229	HCC181	*Chemotherapy*	Sí	2	Binario
230	HCC182	*Rehabilitation*	Sí	2	Binario
231	HCC183	*Screening/Observation/Special Exams*	Sí	2	Binario
232	HCC184	*History of Disease*	Sí	2	Binario
233	COMORB*	*Comorbidity*	Sí	6	T13
234	RXTYPE	*Type of diabetes prescription*	Sí	4	T14

Tabla 4.1 Variables de los datos DxCG.

Las variables marcadas con * son variables derivadas a partir de los datos proporcionados por DxCG. La forma en que se obtienen estas variables se muestra en la siguiente tabla:

Variable	Descripción	Cálculo para obtener la variable
X_1	*Q1 costs normalized to the DXCG estimate*	4 ACOVY_1 / COSTDXCG
X_2	*Q2 costs normalized to the DXCG estimate*	4 ACOVY_2 / COSTDXCG
X_3	*Q3 costs normalized to the DXCG estimate*	4 ACOVY_3 / COSTDXCG
X_4	*Q4 costs normalized to the DXCG estimate*	4 ACOVY_4 / COSTDXCG
SLOPE	*Linear regression slope*	(X_1 + 2 X_2 + 3 X_3 + 4 X_4) / 4 – 2.5 (X_1 + X_2 + X_3 + X_4) / 1.25
OFFSET	*Linear regression offset*	(X_1 + X_2 + X_3 + X_4) / 4 – 2.5 SLOPE
SUDDENNESS	*Surge in costs in the fourth quarter*	4 (X_4 – (4 SLOPE + OFFSET)) / (X_1 + X_2 + X_4 + X_4)
CONCAVITY	*Concavity of the quarterly costs*	2 (X_2 + X_3 – X_1 – X_4) / (X_1 + X_2 + X_4 + X_4)
RMSE	*Root mean squared error of the linear fit to quarterly costs*	RAIZ[((X_ 1 – OFFSET – SLOPE)^2 + (X_2 – OFFSET – 2 SLOPE)^2 + (X_3 – OFFSET – 3 SLOPE)^2 + (X_4 – OFFSET – 4 SLOPE)^2) / 4]
NLR	*Non-linearity of the quarterly costs*	4 RMSE / (X_1+X_2+X_3+X_4)
HOS1	*Number of hospitalization days in Q1*	COVYI_1 / (COVYI / TLOS)
HOS2	*Number of hospitalization days in Q2*	COVYI_2 / (COVYI / TLOS)
HOS3	*Number of hospitalization days in Q3*	COVYI_3 / (COVYI / TLOS)
HOS4	*Number of hospitalization days in Q4*	COVYI_4 / (COVYI / TLOS)

Variable	Descripción	Cálculo para obtener la variable
COSTRI4	*Change in total spending from Q1, Q2, Q3 to Q4*	TOP[ACOVY_4] – TOP[(ACOVY_1+ ACOVY_2 + ACOVY_3) / 3] si el resultado es ≤ 0 se guarda 0
COSTRI43	*Change in total spending from Q1, Q2 to Q3, Q4*	TOP[(ACOVY_3 + ACOVY_4) / 2] - TOP[(ACOVY_1+ ACOVY_2 + ACOVY_3) / 3] si el resultado es ≤ 0 se guarda 0
COSTRII4	*Change in total inpatient spending from Q1, Q2, Q3 to Q4*	TOP[COVYI_4] – TOP[(COVYI_1+ COVYI_2 + COVYI_3) / 3] si el resultado es ≤ 0 se guarda 0
COSTRII43	*Change in total inpatient spending from Q1, Q2 to Q3, Q4*	TOP[(COVYI_3 + COVYI_4) / 2] - TOP[(COVYI_1+ COVYI_2 + COVYI_3) / 3] si el resultado es ≤ 0 se guarda 0
SCORE	*Score function*	COVYO / < COVYO > + ACOVY / <ACOVY> + ACOVY4 / <ACOVY4> + COSTDXCG / <COSTDXCG>, donde <VAR > indica el promedio de la variable VAR en el 0.5% superior
COMORB	*Comorbidity*	Suma de las 184 variables HCC

Tabla 4.2 Variables derivadas.

Los valores usados de las variables de DxCG para obtener las variables derivadas son los valores sin *coarse graining*, es decir los valores originales.

Además de las variables mostradas, en los datos de entrenamiento se contó con la variable a predecir, ACOVYp1, la cual tiene los costos en el siguiente año del paciente correspondiente a cada registro.

En el apéndice B se muestran las descripciones para cada valor de las variables.

4.2 Análisis inicial de los datos

Antes de realizar el *coarse graining* de las variables se analizaron los valores originales de ellas, tanto para todos los datos de 1997, 1998 y 2000, como para los datos que pertenecen a la clase (0.5% superior de los costos del siguiente año). Las principales estadísticas de estos datos se muestran en el apéndice B.

En este análisis se detectaron anomalías en los datos del 2000, principalmente la falta de valores para la variable COVYD_1 (*Q1 total drug spending*) y la falta de las variables TLOS (*Total length of stay*) y NHOSP (*Number of hospitalization*). Debido a la falta de estas dos últimas variables no se pudieron obtener las variables derivadas HOS1, HOS2, HOS3 y HOS4.

Al analizar todos los datos se observó una tendencia en los costos (internos, externos, de medicinas y totales) por trimestre. Los costos internos se refieren a los costos dentro del hospital y los costos externos se refieren a los costos fuera del hospital. En los tres años se observa que los porcentajes de costos en los dos primeros trimestres son relativamente iguales; del segundo al tercer trimestre hay una reducción sustancial de gastos (aproximadamente del 70%) y del tercer al cuarto trimestre hay un aumento sustancial de gastos (aproximadamente del 45%). Los porcentajes obtenidos por

trimestre, para cada año, se obtuvieron con respecto a los costos anuales totales. La tendencia mencionada se puede observar en la figura 4.1.

También se observó que la proporción de los costos internos, externos y de medicinas con respecto al total anual era aproximadamente la misma en los tres años, como se muestra en la figura 4.2.

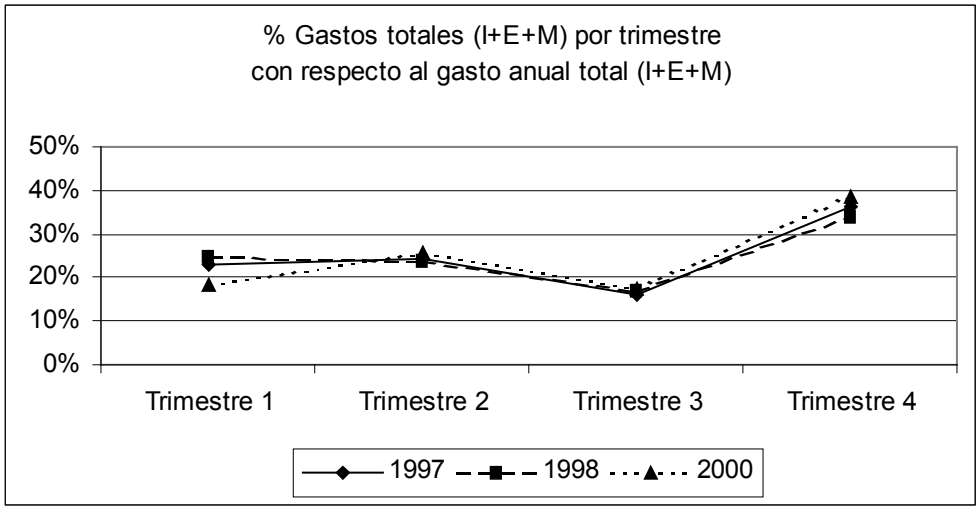

Fig. 4.1 Tendencias de los costos totales durante los años 1997, 1998 y 2000. La figura indica los porcentajes de costos por trimestre, con respecto a los costos anuales (años 1997, 1998 y 2000).

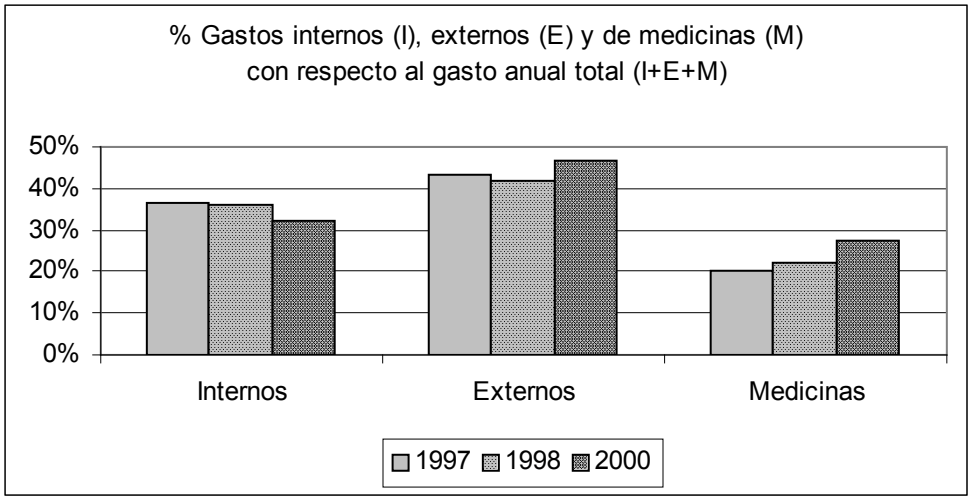

Fig. 4.2 Porcentaje de los costos internos, externos y de medicinas con respecto a los costos totales anuales.

Otro aspecto de interés fue que el promedio de los costos totales del año 2000 fue menor que el promedio de los costos totales de 1997 y 1998. El año que tuvo mayor promedio de costos totales fue 1998, como se puede apreciar en la figura 4.3. Hay que recordar que los grupos de pacientes de los tres años fueron diferentes.

Las gráficas de dispersión de los años 1997 – 1998, 1997 – 2000 y 1998 – 2000 muestran que los datos son estadísticamente muy similares entre los tres años. El eje X de estas gráficas corresponde a las medias de un conjunto de variables en un año. El eje Y corresponde a las medias de las mismas variables, pero en otro año. Las figuras 4.4 a 4.6 muestran la dispersión de las variables de costos de los años 1997 – 1998, 1997 – 2000 y 1998 – 2000, respectivamente. Las figuras 4.7 a 4.9 muestran la dispersión de las variables HCC de los años 1997 – 1998, 1997 – 2000 y 1998 – 2000, respectivamente.

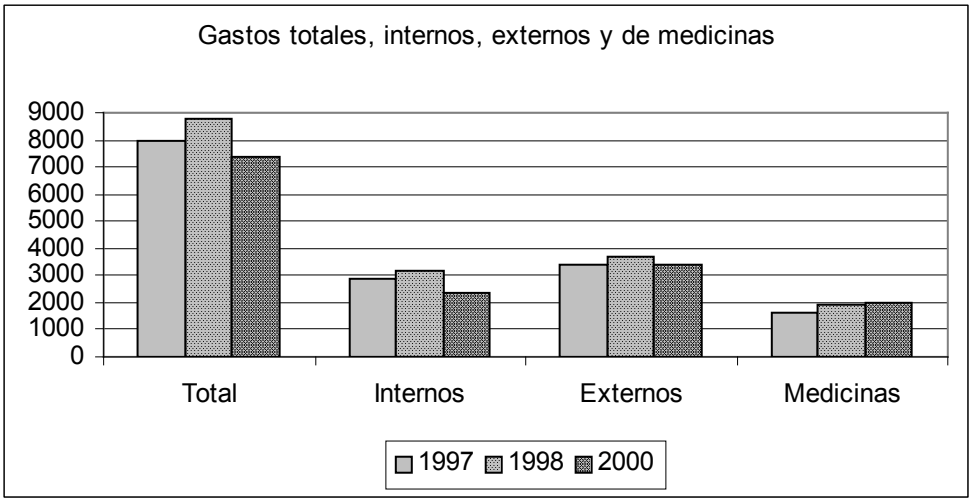

Fig. 4.3 Costos totales, internos, externos y de medicinas para los años 1997, 1998 y 2000. Los costos están en dólares.

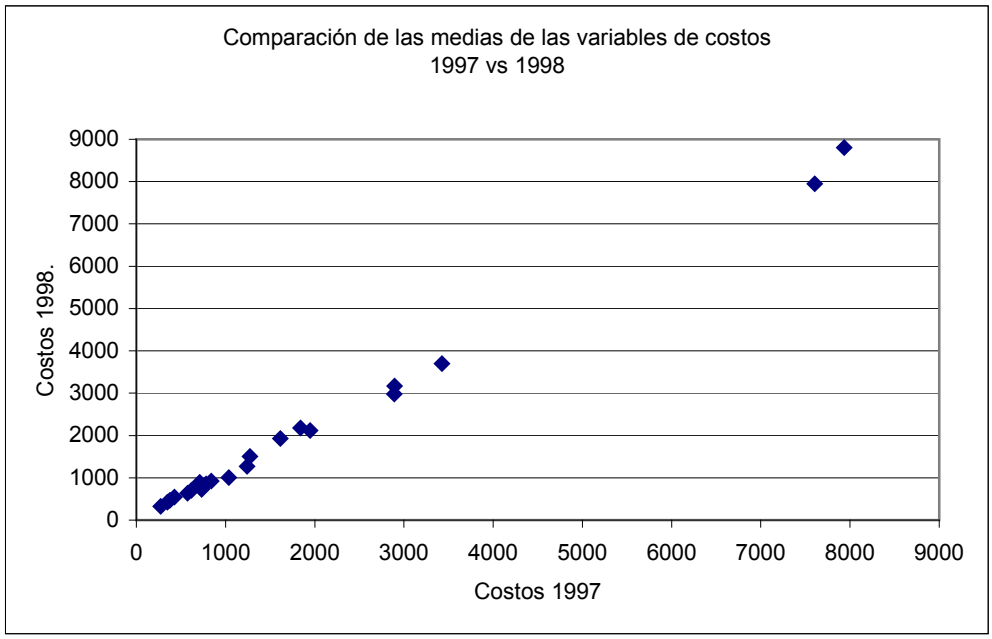

Fig. 4.4 Comparación de las medias de las variables de costos de 1997 versus 1998. Los costos están en dólares.

En las figuras de costos se observa que en los años 1997 – 1998, prácticamente no hay diferencias entre las medias de las variables, salvo que se observa que en general, los costos de 1998 fueron mayores a los costos de 1997. Al comparar 1997 y 1998 contra el

2000, se observa que hay unos puntos que sobresalen, indicando diferencias entre los distintos años. Tanto en la figura 4.5, como en la figura 4.6 estos puntos corresponden a las variables COVYD_1 (*Q1 total drug spending*), ACOVY_1 (*Q1 total spending I+O+D*), COVYI (*Total inpatient spending*) y ACOVY (*Total spending I+O+D*). Las diferencias con la variable COVYD_1 se explican porque no se tienen valores de esta variable en el año 2000. Las diferencias en las otras variables indican que el promedio de costos totales del primer trimestre de 1997 y del primer trimestre de 1998 fue menor al promedio de los costos totales del primer trimestre del 2000, y que el promedio de costos internos de 1997 y 1998 fue menor al promedio de costos internos del 2000. Estas diferencias repercuten en los promedios totales de costos de los años (variable ACOVY). También se puede observar que los costos promedio del 2000 fueron menores que los de 1998, como ya se había mostrado en la figura 4.3.

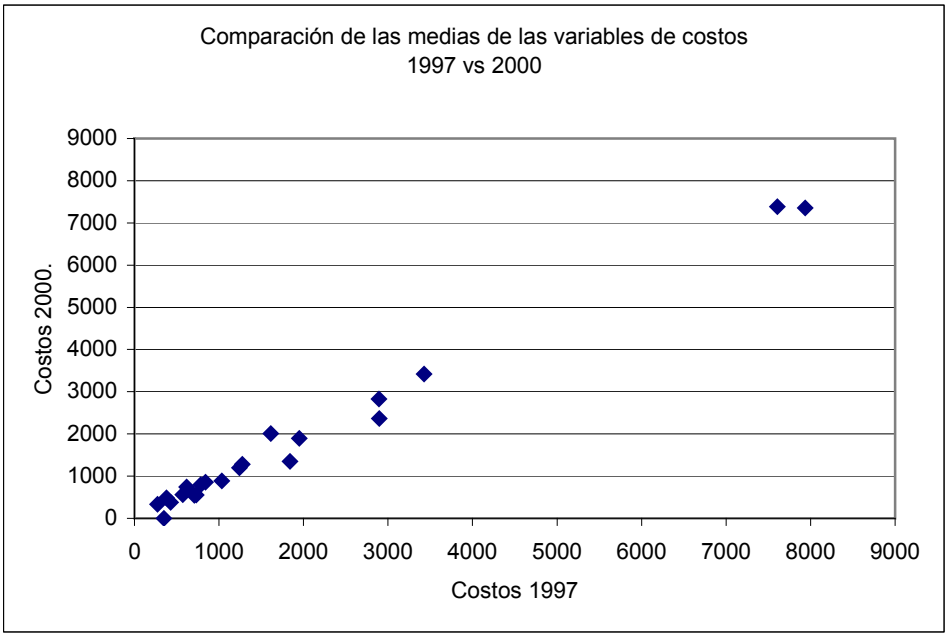

Fig. 4.5 Comparación de las medias de las variables de costos de 1997 versus 2000.

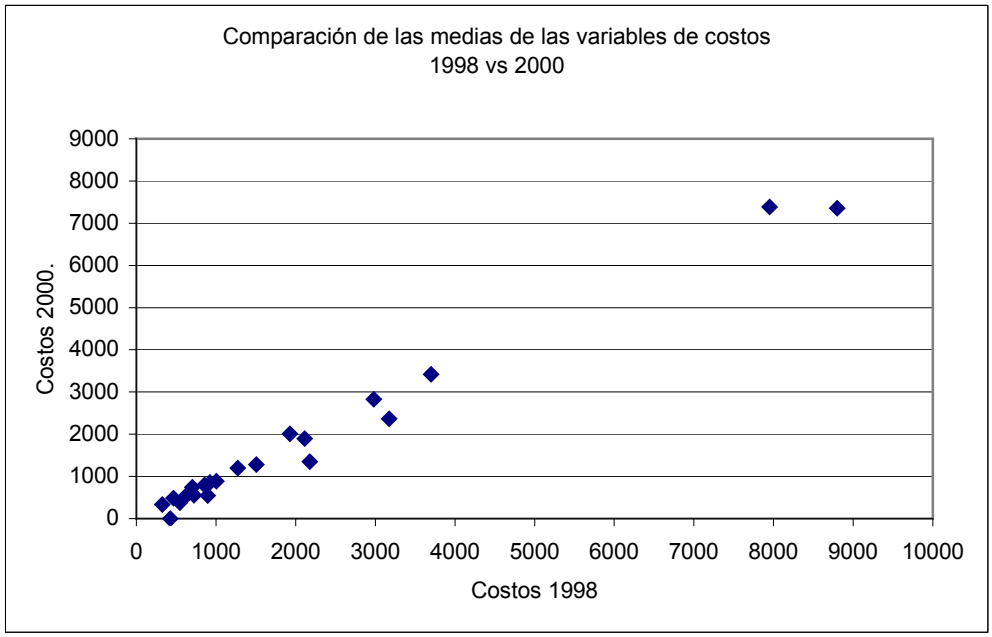

Fig. 4.6 Comparación de las medias de las variables de costos de 1998 versus 2000.

Con respecto a las variables HCC tampoco se encontraron diferencias entre las variables de 1997 y 1998 y sí se encontraron algunas diferencias con respecto a los datos del 2000. Las diferencias se dan principalmente en las variables HCC014 (*Benign Neoplasms of Skin, Breast, Eye*), HCC167 (*Minor Symptoms, Signs, Findings*) y HCC043 (*Other Musculoskeletal and Connective Tissue Disorders*). El punto casi cercano a (1,1) corresponde a la variable HCC019 (*Diabetes with No or Unspecified Complications*), lo que indica que la mayoría de los pacientes presentan una diabetes sin complicaciones o con complicaciones no especificadas.

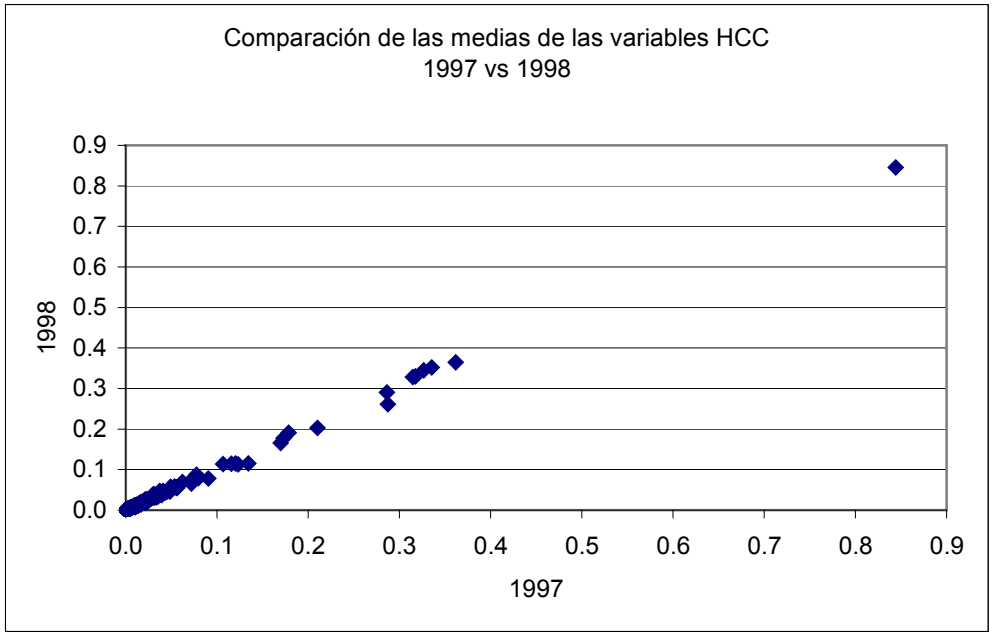

Fig. 4.7 Comparación de las medias de las variables HCC de 1997 versus 1998.

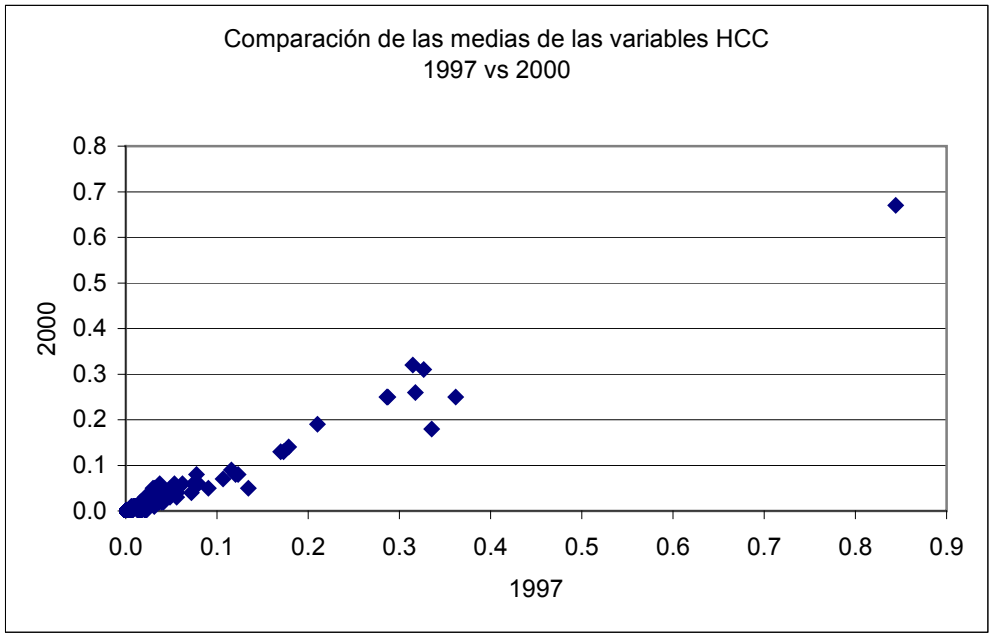

Fig. 4.8 Comparación de las medias de las variables HCC de 1997 versus 2000.

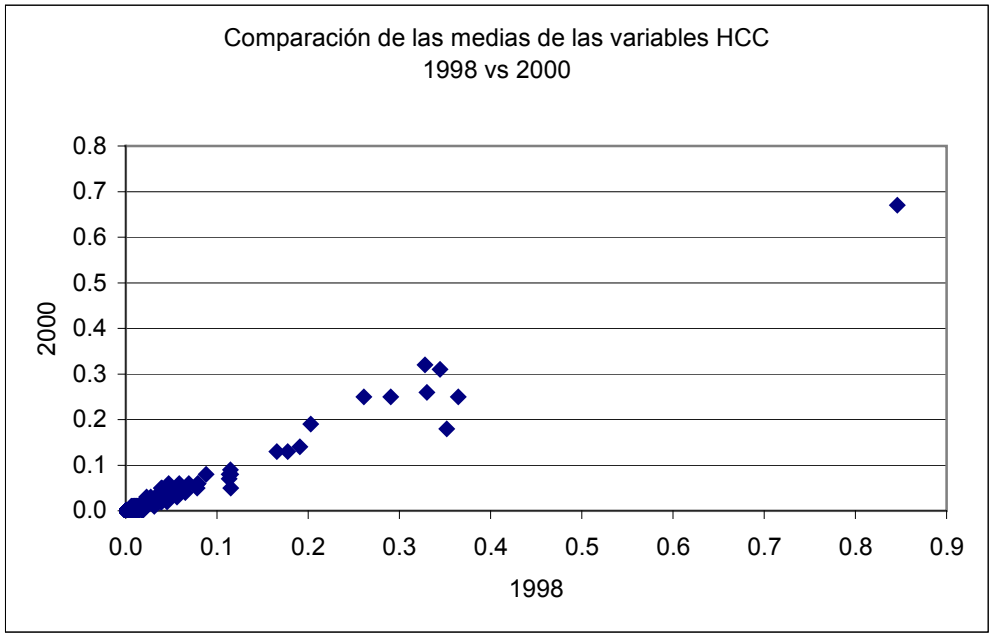

Fig. 4.9 Comparación de las medias de las variables HCC de 1998 versus 2000.

También se obtuvieron las mismas estadísticas por año 1997 y 1998 (mínimo, máximo, media y desviación estándar) para los datos que pertenecen a la clase. Esto se hizo para encontrar las variables que dieran una buena discriminación de la clase.

Comparando estas estadísticas se encontró que en los datos que pertenecen a la clase, los porcentajes de costos internos y externos aumentaron y los porcentajes de costos de medicinas disminuyeron. Este comportamiento se observó en los datos de ambos años como se puede apreciar en las figuras 4.10 y 4.11. Los porcentajes de los costos internos,

externos y de medicinas se obtuvieron con respecto a los costos totales del año analizado.

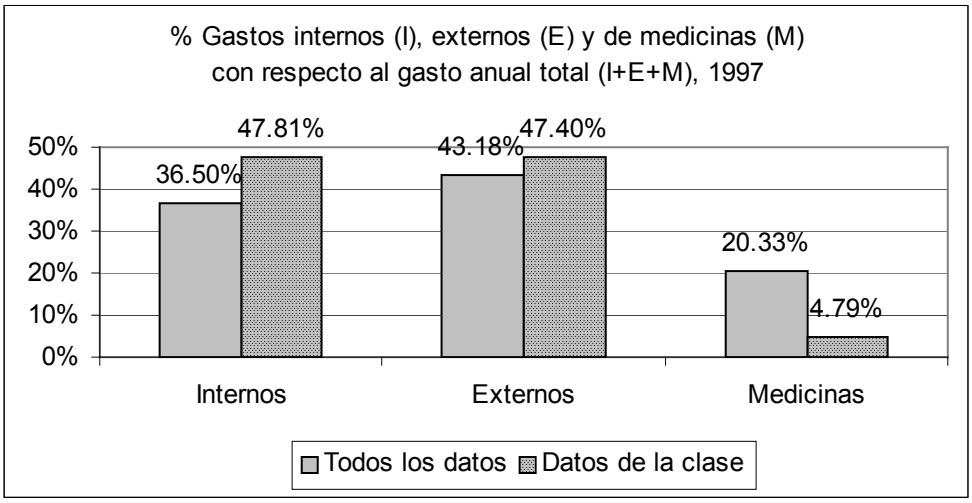

Fig. 4.10 Comparación de los costos de 1997 entre todos los datos y los datos de la clase.

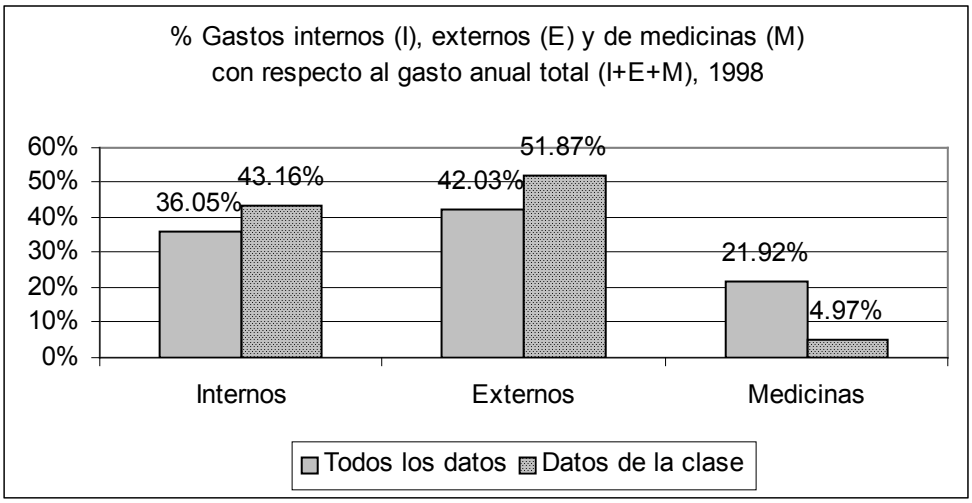

Fig. 4.11 Comparación de los costos de 1998 entre todos los datos y los datos de la clase.

Comparando todos los datos contra los datos de la clase se encontró que el promedio de los costos totales aumentó más de diez veces en los datos de la clase. El tipo de costos que más aumentó en los datos de la clase fueron los costos internos: los costos internos promedio de la clase fueron aproximadamente 15 veces más que los costos internos promedio de todos los datos. El aumento de costos de medicinas en los datos de la clase fue solamente de un factor de aproximadamente 2.5. Esto se puede apreciar en la gráfica 4.12.

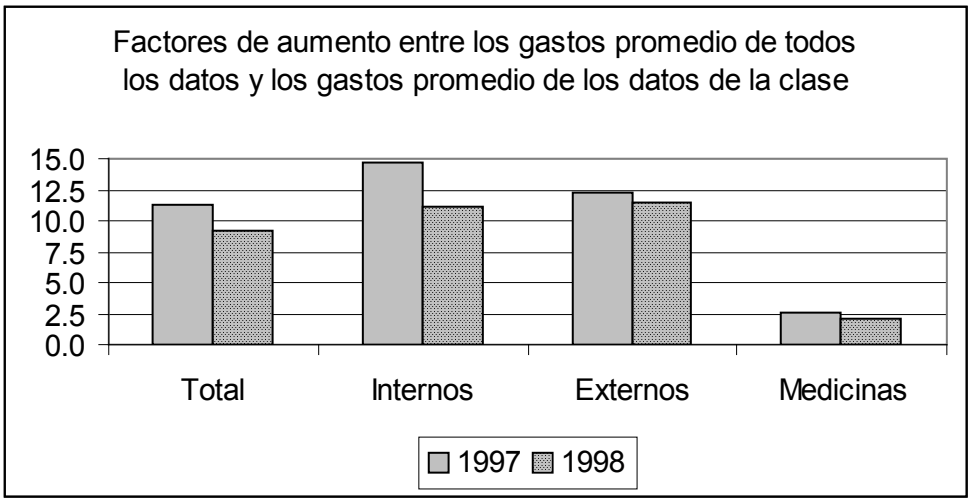

Fig. 4.12 Comparación de los costos de 1998 entre todos los datos y los datos de la clase.

También se observó que en los datos de la clase los porcentajes de los costos externos disminuyeron en el primer trimestre y aumentaron en el cuarto trimestre como se puede apreciar en las siguientes dos gráficas.

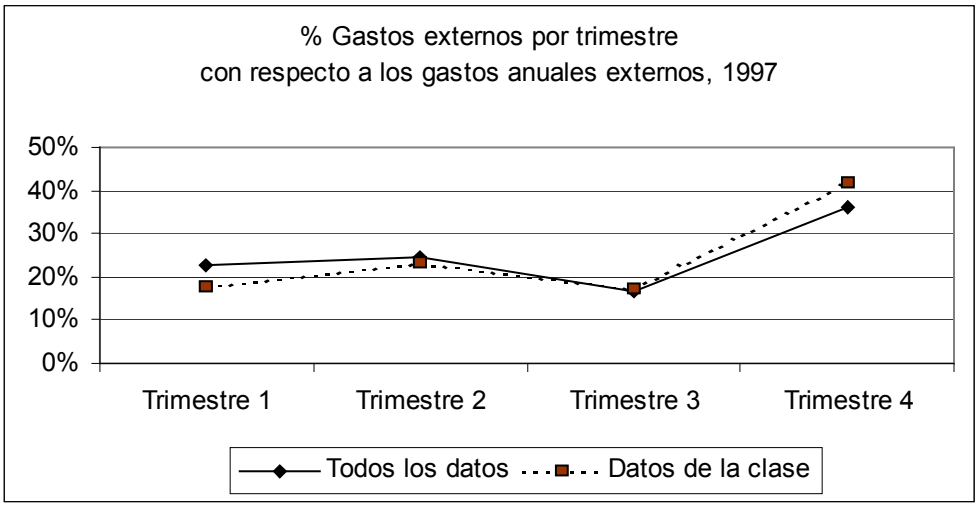

Fig. 4.13 Comparación de los costos externos de 1997 entre todos los datos y los datos de la clase.

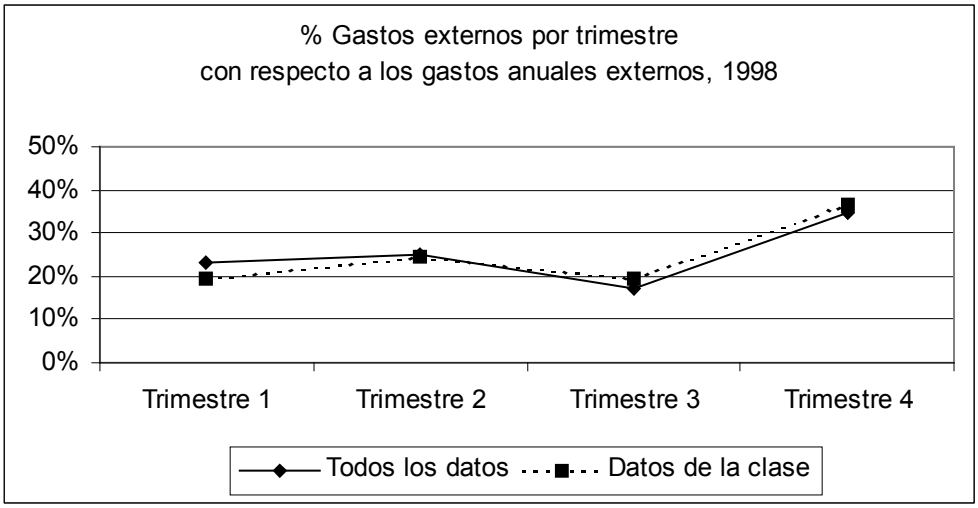

Fig. 4.14 Comparación de los costos externos de 1998 entre todos los datos y los datos de la clase.

Otras dos variables que tuvieron grandes diferencias fueron COMORB y COSTDXCG. La variable COMORB indica el número de variables HCC que tienen valor 1 (condición médica presente). La variable COSTDXCG es el costo estimado por DxCG para el año siguiente de acuerdo a su modelo de regresión lineal, el cual solo toma en cuenta la condición médica del paciente. Los pacientes que están en el 0.5% superior más costoso tienen un promedio de comorbidad mayor que el de todos los pacientes. Este comportamiento es natural, ya que mientras más trastornos o enfermedades tenga un paciente, más tratamientos requiere y por lo tanto sus costos aumentan. Dado que la variable COSTDXCG es la predicción de los costos del paciente en el siguiente año, entonces es natural que sea un buen discriminador de la clase. Debido a esto, es importante incluir a la variable COSTDXCG en el modelo de predicción a construir. Este es el enfoque multiperspectiva que se desea resaltar en el presente trabajo: para la resolución de un problema de minería de datos hay que tomar en cuenta las opiniones de diversos agentes e integrarlas para obtener mejores resultados, en este caso se está integrando en los datos la opinión de un modelo de regresión lineal.
Con respecto a las variables HCC, existen varias condiciones médicas que están más presentes en los casos de pacientes que pertenecen a la clase. En la siguiente sección se verá cuales son las variables HCC que discriminan mejor la clase.

4.2.1 Funciones ε' y ε

Después del análisis inicial de los datos se obtuvieron los valores de ε' y ε para encontrar las variables más predictivas.

Para poder calcular los valores de ε y ε' se hizo el *coarse graining* de las variables. En el apéndice B se muestran las estadísticas principales (mínimo, máximo, media y desviación estándar) de las variables con *coarse graining*.

En esta sección sólo se muestran los diez resultados más altos y los diez resultados más bajos de ε' y ε para los años 1997 y 1998. En el apéndice B se muestran todos los valores de ε' y ε calculados.

N	Variable	Descripción	ε'
1	COSTDXCG	*Cost estimate by DXCG model*	17.40
2	ACOVY	*Total spending I+O+D*	17.31
3	COVYO	*Total outpatient spending*	15.38
4	ACOVY_4	*Q4 total spending I+O+D*	14.89
5	COMORB	*Comorbidity*	14.69
6	COVYO_4	*Q4 total outpatient spending*	13.53
7	ACOVY_2	*Q2 total spending I+O+D*	12.88
8	ACOVY_3	*Q3 total spending I+O+D*	12.17
9	NHOSP	*Number of hospitalizations*	11.25
10	COVYO_2	*Q2 total outpatient spending*	11.13

Tabla 4.3 Valores más altos de ε' en los datos de 1997.

En la tabla 4.3 resalta el hecho de que las variables que más discriminan a la clase son las variables relacionadas con los costos. Sobre todo los costos totales y los costos externos. En la figura 4.12 se observó que los costos internos aumentaron en un factor de 15 en los datos de la clase. La razón de que los costos internos no aparezcan en los diez valores más altos de ε' y en cambio sí aparezcan los costos externos quizá se deba a que los costos externos tienen más presencia en los costos totales como se puede apreciar en la figura 4.2. Con respecto a la condición médica del paciente, las únicas dos variables que aparecen son: COMORB (*Comorbity*): entre más trastornos o enfermedades tenga un paciente es más probable que se incrementen sus costos y NHOSP (*Number of hospitalizations*) con el mismo efecto.

N	Variable	Valor	Descripción	ε
1	COVYO	6	*Total outpatient spending: Top 0.5%*	48.65
2	COVYO_4	6	*Q4 total outpatient spending: Top 0.5%*	40.40
3	ACOVY	6	*Total spending I+O+D: Top 0.5%*	38.04
4	HCC131	1	*Renal Failure: Present*	37.00
5	COVYO_2	6	*Q2 total outpatient spending: Top 0.5%*	34.51
6	COSTDXCG	6	*Cost estimate by DXCG model: Top 0.5%*	33.33
7	COVYO_3	6	*Q3 total outpatient spending: Top 0.5%*	32.15
8	ACOVY_4	6	*Q4 total spending I+O+D: Top 0.5%*	30.97
9	COVYO_1	6	*Q1 total outpatient spending: Top 0.5%*	28.61
10	HCC130	1	*Dialysis Status: Present*	27.44

Tabla 4.4 Valores más altos de ε en los datos de 1997.

En la tabla 4.4 aparecen casi las mismas variables que en la tabla 4.3. La información adicional que se da en esta tabla es que estas variables repetidas tienen un valor que corresponde al 0.5% superior de la variables (el valor 6 corresponde al 0.5% superior de la variable), es decir el 0.5% de los pacientes más costosos. También aparecen otras dos variables de condición médica: la presencia de fallas renales y diálisis.

N	Variable	Descripción	ε'
1	HCC146	*Uncompleted Pregnancy With Complications*	-22.26
2	HCC144	*Completed Pregnancy With Complications*	-20.19
3	HCC155	*Major Head Injury*	-11.43

N	Variable	Descripción	ε'
4	HCC098	*Cerebral Atherosclerosis and Aneurysm*	-10.31
5	HCC143	*Completed Pregnancy With Major Complications*	-10.27
6	HCC049	*Dementia*	-9.61
7	HCC072	*Multiple Sclerosis*	-9.01
8	HCC066	*Attention Deficit Disorder*	-8.61
9	HCC054	*Schizophrenia*	-8.38
10	HCC137	*Female Infertility*	-8.38

Tabla 4.5 Valores más bajos de ε' en los datos de 1997.

La tabla 4.5 muestra que estas condiciones médicas no repercuten mucho en los costos de los pacientes. En esta tabla no aparecen variables relacionadas con los costos.

N	Variable	Valor	Descripción	ε
1	ACOVY	1	*Total spending I+O+D: Bottom 80%*	-8.47
2	COSTDXCG	1	*Cost estimate by DXCG model: Bottom 80%*	-8.19
3	COVYO	1	*Total outpatient spending: Bottom 80%*	-8.10
4	ACOVY_4	1	*Q4 total spending I+O+D: Bottom 80%*	-7.73
5	ACOVY_2	1	*Q2 total spending I+O+D: Bottom 80%*	-7.35
6	COVYO_4	1	*Q4 total outpatient spending: Bottom 80%*	-7.26
7	TLOS	0	*Total length of stay: 0 days*	-6.78
8	NHOSP	0	*Number of hospitalizations: 0 hospitalization*	-6.78
9	ACOVY_3	1	*Q3 total spending I+O+D: Bottom 80%*	-6.70
10	COVYO_2	1	*Q2 total outpatient spending: Bottom 80%*	-6.61

Tabla 4.6 Valores más bajos de ε en los datos de 1997.

La tabla 4.6 muestra que las variables de costos con valores que indican los costos menos altos (valor 1) discriminan bien a la clase en el sentido de que los pacientes con dichos valores son muy poco probables de pertenecer a la clase. Otras variables relacionadas con la condición médica del paciente son el número de hospitalizaciones y el número de días en el hospital.

N	Variable	Descripción	ε'
1	COSTDXCG	*Cost estimate by DXCG model*	21.10
2	ACOVY	*Total spending I+O+D*	18.65
3	COMORB	*Comorbidity*	17.02
4	COVYO	*Total outpatient spending*	15.92
5	ACOVY_4	*Q4 total spending I+O+D*	15.57
6	ACOVY_3	*Q3 total spending I+O+D*	14.41
7	COVYO_4	*Q4 total outpatient spending*	13.87
8	ACOVY_2	*Q2 total spending I+O+D*	13.80
9	NHOSP	*Number of hospitalizations*	13.11
10	COVYO_3	*Q3 total outpatient spending*	12.98

Tabla 4.7 Valores más altos de ε' en los datos de 1998.

Las variables de la tabla 4.7 son las mismas que para el año 1997. La única diferencia es que en 1998 el valor de ε' de la variable COVYO_3 fue mayor al de COVYO_2 del mismo año (a diferencia de 1997 donde aparece COVYO_2 y no aparece COVYO_3).

N	Variable	Valor	Descripción	ε
1	HCC131	1	*Renal Failure: Present*	43.07
2	COVYO	6	*Total outpatient spending: Top 0.5%*	42.83
3	COVYO_4	6	*Q4 total outpatient spending: Top 0.5%*	40.79
4	COVYO_2	6	*Q2 total outpatient spending: Top 0.5%*	37.73
5	COVYO_3	6	*Q3 total outpatient spending: Top 0.5%*	37.73
6	ACOVY	6	*Total spending I+O+D: Top 0.5%*	36.71
7	COSTDXCG	6	*Cost estimate by DXCG model: Top 0.5%*	34.68
8	HCC130	1	*Dialysis Status: Present*	32.93
9	NHOSP	3	*Number of hospitalizations: >= 4 hospitalization*	29.80
10	COVYO_1	6	*Q1 total outpatient spending: Top 0.5%*	29.58

Tabla 4.8 Valores más altos de ε en los datos de 1998.

Nuevamente las variables de la tabla 4.8 son las mismas que para el año 1997. La única diferencia es que en 1998 el valor de ε de la variable NHOSP=3 fue mayor al de ACOVY_4=6 del mismo año (a diferencia de 1997 donde aparece ACOVY_4=6 y no aparece NHOSP=3).

N	Variable	Descripción	ε'
1	HCC125	*Significant Ear, Nose, and Throat Disorders*	-14.18
2	HCC137	*Female Infertility*	-10.93
3	HCC054	*Schizophrenia*	-9.71
4	HCC066	*Attention Deficit Disorder*	-9.44
5	HCC160	*Internal Injuries*	-8.13
6	HCC117	*Major Eye Infections/Inflammations*	-7.88
7	HCC073	*Parkinson's and Huntington's Diseases*	-6.79
8	HCC142	*Miscarriage/Abortion*	-6.79
9	HCC101	*Diplegia (Upper), Monoplegia, and Other Paralytic Syndromes*	-6.64

N	Variable	Descripción	ε'
10	HCC004	*Tuberculosis*	-6.56

Tabla 4.9 Valores más bajos de ε' en los datos de 1998.

La tabla 4.9 sí muestra diferencias con respecto a las variables con los valores más bajos de ε' para 1997, pero lo que tienen en común es que solamente son variables de condición médica, no hay variables relacionadas con los costos.

N	Variable	Valor	Descripción	ε
1	COSTDXCG	1	*Cost estimate by DXCG model: Bottom 80%*	-10.24
2	ACOVY	1	*Total spending I+O+D: Bottom 80%*	-9.91
3	COVYO	1	*Total outpatient spending: Bottom 80%*	-8.87
4	ACOVY_4	1	*Q4 total spending I+O+D: Bottom 80%*	-8.55
5	ACOVY_3	1	*Q3 total spending I+O+D: Bottom 80%*	-8.30
6	COVYO_4	1	*Q4 total outpatient spending: Bottom 80%*	-7.90
7	COVYO_3	1	*Q3 total outpatient spending: Bottom 80%*	-7.66
8	TLOS	0	*Total length of stay: 0 days*	-7.61
9	NHOSP	0	*Number of hospitalizations: 0 hospitalization*	-7.61
10	ACOVY_2	1	*Q2 total spending I+O+D: Bottom 80%*	-7.58

Tabla 4.10 Valores más bajos de ε en los datos de 1998.

Las variables de la tabla 4.10 son las mismas que para el año 1997. La única diferencia es que en 1998 el valor de ε' de la variable COVYO_3=1 fue mayor al de COVYO_2=1 del mismo año (a diferencia de 1997 donde aparece COVYO_2=1 y no aparece COVYO_3=1).

De las tablas mostradas en esta sección se aprecia la importancia que tienen las variables que corresponden a los costos para discriminar a la clase, principalmente los costos totales y externos. Dentro de estos costos, los costos correspondientes al cuarto trimestre son los más relevantes. Otras variables de relevancia fueron COSTDXCG, COMORB, TLOS y NHOSP.

4.2.2 Variable *SCORE*

En los perfiles de los años 1997 y 1998 se encontró que las variables más predictivas, de acuerdo con ε', fueron COSTDXCG, ACOVY, COVYO y ACOVY_4. Con estas variables se construyó una nueva variable mediante la relación dada en la tabla 4.2:

$$\text{SCORE} = \frac{\text{COVYO}}{\langle \text{COVYO} \rangle} + \frac{\text{ACOVY}}{\langle \text{ACOVY} \rangle} + \frac{\text{ACOVY4}}{\langle \text{ACOVY4} \rangle} + \frac{\text{COSTDXCG}}{\langle \text{COSTDXCG} \rangle} \qquad \textbf{(4.1)}$$

donde:

$\langle VARIABLE \rangle$ indica el promedio de la variable en el 0.5% superior de dicha variable.

Mientras mayor sea esta variable indica que se tiene un valor arriba del valor promedio en el 0.5% superior, en una o más de las principales variables.

Los valores de ε' y ε de la nueva variable SCORE se muestran en la siguiente tabla.

Año	Variable	Descripción	ε'	ε
1997	SCORE	*Score function*	18.36	47.47
1998	SCORE	*Score function*	20.79	42.83

Tabla 4.11 Valores de ε' y ε' de la variable SCORE en los datos de 1997 y 1998.

Comparando con los valores de las tablas 4.3, 4.4, 4.7 y 4.8 se observa que la variable SCORE se encuentra entre las tres primeras variables con los valores más altos de ε' y ε.

Esta nueva variable fue la más predictiva de todas las variables como se verá en la siguiente sección. El hecho de que la variable SCORE fuera muy predictiva confirma la importancia de adoptar un enfoque multiperspectiva mediante el cual se combinan diferentes opiniones, en este caso la opinión de cuatro diferentes variables, una de ellas conteniendo información de un modelo de regresión lineal.

4.3 Predicción

Debido a la gran dimensión del espacio de características se usó un algoritmo genético para descubrir los clasificadores que fueran más predictivos, por la misma razón, se decidió que el algoritmo genético buscara los mejores clasificadores sobre todas las variables, en lugar de limitar la búsqueda en un número reducido de variables. En este contexto, un clasificador es una cadena que indica valores específicos de las variables. Estos clasificadores se usan para seleccionar un subconjunto de los datos que cumplen con los valores especificados por el clasificador. En la siguiente sección se explicará con más detalle en qué consisten los clasificadores encontrados mediante un algoritmo genético.

4.3.1 Obtención de clasificadores mediante un algoritmo genético

Para determinar cuales eran los clasificadores más predictivos se usaron dos funciones de aptitud: $p(C|\boldsymbol{X})$ y $\varepsilon(\boldsymbol{X})$, las cuales se explican con mayor detalle en la siguiente sección. Esta función de aptitud es la función a maximizar en el algoritmo genético. Dado que los clasificadores obtenidos son aquellos que tienen mayor $p(C|\boldsymbol{X})$, o $\varepsilon(\boldsymbol{X})$, los individuos seleccionados con estos clasificadores tienen mayor probabilidad de pertenecer a la clase.

Un clasificador consiste en una cadena de "bits". Cada bit corresponde a una variable y puede tener N+1 valores: los N posibles valores de las variables más un valor especial, el "*", que sirve para indicar que la variable puede tomar cualquier valor. Para dejar esto más claro considérese que se desean obtener los clasificadores para un conjunto de datos con dos variables binarias. En este caso, se pueden tener hasta nueve clasificadores: 00, 01, 0*, 10, 11, 1*, *0, *1 y **. El clasificador 00 selecciona a todos los individuos que tienen 0 en sus dos variables, el clasificador 01 selecciona a los individuos que tienen 0 en su primera variable y 1 en su segunda variable, el clasificador 0* selecciona a los individuos que tienen 0 en su primer variable y 0 ó 1 en su segunda variable, y así sucesivamente.

La importancia del valor especial "*" se resalta en los casos en que se tienen muchas variables y pocos datos. Para explicar esto considérese que los datos tienen veinte

variables y que todas las variables son binarias. En este caso, se pueden tener hasta 1,048,576 clasificadores sin usar el caracter especial. Si se cuenta con menos de 1,048,576 registros (lo cual es muy probable), entonces habrá algunos clasificadores que no correspondan a ningún individuo de los datos. Este problema se agrava cuando el número de variables y su cardinalidad aumenta. En estos casos, lo mejor es usar el valor *comodín* "*": si un bit tiene el valor "*" significa que la variable correspondiente puede tener cualquier valor. Por ejemplo, el clasificador representado por la cadena "1********0*********1" selecciona a los individuos que tienen 1 en la primera variable, 0 en la décima variable, 1 en la vigésima variable y el resto de las variables pueden tener 0 ó 1. De esta manera es muy probable que los clasificadores encontrados disparen a más de un individuo y por lo tanto sean más significativos estadísticamente.

La población inicial del algoritmo genético fue un conjunto de clasificadores determinados aleatoriamente y usando una probabilidad alta para asignar el carácter especial "*" en cualquier variable. Esto se hizo con la finalidad de que un clasificador no tuviera más de cuatro o cinco bits con valores específicos, ya que la dimensión del espacio es muy grande y el conjunto de datos es reducido.

Se realizaron varias pruebas para determinar los mejores parámetros del algoritmo genético (número de generaciones, tamaño de la población, probabilidad de mutación y probabilidad de cruzamiento). Estos parámetros se muestran en la sección 4.4.1. Se usó cruzamiento de un solo punto, así como selección tipo ruleta. También se empleó elitismo con memoria, ya que durante el entrenamiento de una corrida del algoritmo genético se conservó la lista de los 100 mejores clasificadores. Para determinar los mejores parámetros del algoritmo genético se midió el desempeño en los datos de entrenamiento. El conjunto de datos de entrenamiento usado fueron los datos de 1997.

Con los parámetros determinados en el paso anterior se realizaron veinte multi-corridas. Una multi-corrida del algoritmo genético consiste de varias corridas internas. Estas corridas internas permiten diversificar la búsqueda en el espacio de características. Por razones de tiempo se eligieron 10 corridas internas. Al final del entrenamiento se obtuvo una lista final con los mejores cien clasificadores. Estos clasificadores se usaron para seleccionar el 0.5% superior de individuos de los datos de entrenamiento. Con los clasificadores obtenidos en el entrenamiento se seleccionó el 0.5% superior en los datos de prueba.

4.3.2 Funciones de aptitud del algoritmo genético

Se consideraron dos funciones de aptitud:

1. Probabilidad $p(C|\boldsymbol{X})$
2. $\varepsilon(\boldsymbol{X})$.

En ambas funciones la clase es el 0.5% superior de pacientes más costosos y $\boldsymbol{X}$ se refiere a un clasificador en particular como los definidos en la sección 4.3.1. Mediante el algoritmo genético se buscaron los clasificadores que maximizaran estas funciones.

En el capítulo anterior se vio que mediante la función $p(C|\boldsymbol{X})$ se encuentran pocos individuos que corresponden a un vector $\boldsymbol{X}$ en particular. Esto tiene la desventaja de que no se tiene suficiente confiabilidad estadística, sin embargo la ventaja es que se puede tener mayor precisión en la predicción. Por otra parte, la función ε proporciona mayor

confiabilidad estadística, pero el problema es que puede resaltar clasificadores no muy predictivos que están asociados a un número grande de pacientes. Debido a las ventajas de ambas funciones de aptitud, se decidió considerar ambas funciones y buscar una manera de aminorar sus desventajas.

4.3.3 El problema de los clasificadores redundantes

El uso del comodín "*" en los clasificadores implica que habrá unos clasificadores que sean redundantes. Considérese el caso de los clasificadores para tres variables binarias, En este caso se tienen nueve posibles clasificadores. Debido al uso del caracter especial "*" algunos clasificadores serán redundantes. Por ejemplo, el clasificador 11* comprende a los clasificadores 110 y 111. El problema que surge con la redundancia es que puede originar que los clasificadores se ordenen incorrectamente mediante su aptitud. Supóngase que las aptitudes de los clasificadores 101, 11* y 110 están ordenados de la siguiente manera: $f_{111} > f_{11^*} > f_{10^*}$. Como se puede apreciar, los dos primeros clasificadores son redundantes. Si $f_{10^*} > f_{110}$, entonces el ranqueo 111, 11*, 10* es incorrecto en el sentido de que f_{11^*} es mayor a f_{10^*} debido al clasificador 111. Si se elimina esta redundancia y se dejan los clasificadores 111, 110 y 10* entonces el orden apropiado sería $f_{111} > f_{110} > f_{10^*}$. Una manera de aminorar los problemas de la redundancia es ordenando los clasificadores usando un nuevo criterio de ordenamiento multiperspectiva como se verá en la siguiente sección.

4.3.4 El enfoque multiperspectiva

Mediante el algoritmo genético se obtiene la lista de los clasificadores más predictivos. Ya sea usando la función de aptitud $p(C|\boldsymbol{X})$ o $\varepsilon(\boldsymbol{X})$. Los clasificadores están ordenados inicialmente por su aptitud y la manera en que se elige a los individuos más probables de pertenecer al 0.5% superior más costoso es recorriendo la lista de clasificadores, de mayor a menor aptitud, e ir seleccionando los individuos que corresponden a cada clasificador hasta completar el 0.5% deseado.

Sin embargo, esta no es la única manera de ordenar los clasificadores o de seleccionar los individuos. Se pueden seguir otros criterios que mejoran los resultados de la predicción. A continuación se mencionan otros criterios que se pueden seguir. Los criterios presentados tienen en común que a cada individuo se le asigna una puntuación y al final se eligen los individuos con mayor puntuación.

1. **Ganador (G)**. En este criterio la puntuación que se le asigna a un individuo es la mayor aptitud de todos los clasificadores que dispara el individuo. La función de aptitud empleada para asignar la puntuación es la misma función que se maximiza en el algoritmo genético. En el caso de empates de aptitudes entre dos o más individuos, el desempate se realiza dando preferencia al individuo que dispara un mayor número de clasificadores.

2. **Ganador con re-ranqueo (RR)**. En este criterio la puntuación que se le asigna a un individuo también es la aptitud más grande de todos los clasificadores que dispara el individuo, pero a diferencia del criterio anterior, la función de aptitud empleada para asignar la puntuación no es la función a maximizar con el algoritmo genético, sino la otra función de aptitud. Por ejemplo, si con el algoritmo genético se encontraron los clasificadores que maximizan la función $\varepsilon(\boldsymbol{X})$, para el re-ranqueo se usa $p(C|\boldsymbol{X})$ y viceversa. En caso de empate de

aptitudes entre individuos se da preferencia al individuo que dispara un mayor número de clasificadores. Mediante este criterio se busca reducir las desventajas de las funciones de aptitud mencionadas en la sección 4.3.2.

3. **Aptitud promedio (AP)**. En este criterio la puntuación de un individuo se asigna promediando la aptitud de todos los clasificadores que dispara el individuo. En caso de empate de aptitudes entre individuos se da preferencia al individuo que dispara un mayor número de clasificadores.

4. **Número de correspondencias (NC)**. En este criterio la puntuación que se le asigna a un individuo es el número de clasificadores que dispara el individuo. En caso de empate entre individuos se da preferencia al individuo que tenga el clasificador con más alta aptitud.

5. **Ganador con re-evaluación (RE)**. En este criterio, la lista final de clasificadores se re-ranquea de la siguiente manera: en cada iteración se selecciona el clasificador con la mejor aptitud, se remueven los individuos asociados con dicho clasificador y se recalcula la aptitud de los clasificadores restantes tomando en cuenta los individuos que quedan. Se repiten las iteraciones hasta seleccionar todos los clasificadores, remover todos los individuos, o llegar al número máximo de re-evaluaciones prestablecido. Al final se obtiene la misma lista de clasificadores pero ordenada según como se iban obteniendo las mejores aptitudes durante la re-evaluación. Se seleccionan los individuos mediante el criterio uno, pero usando el nuevo orden de los clasificadores en lugar de la aptitud.

6. **Ganador con re-ranqueo y re-evaluación (RR_RE)**. Este criterio es igual al criterio anterior, pero en lugar de ir quitando los clasificadores con más alta aptitud, se van quitando los clasificadores con la otra función más alta: $p(C|\boldsymbol{X})$ en caso de que la función de aptitud haya sido $\varepsilon(\boldsymbol{X})$ y viceversa. Se seleccionan los individuos mediante el criterio uno, pero usando el nuevo orden de los clasificadores en lugar de la aptitud.

En los criterios "ganador con re-evaluación" y "ganador con re-ranqueo y re-evaluación" se aminora de manera directa el problema de los clasificadores redundantes, ya que la re-evaluación se van removiendo los individuos que corresponden a los clasificadores que se van recorriendo. En el resto de los criterios también se reduce el problema de redundancia de los clasificadores pero de una manera menos directa, al desempatar los individuos mediante el número de clasificadores que dispara cada individuo.

4.3.5 *Benchmarks* de predicción

Para medir el desempeño de la predicción, se usaron dos *benchmarks*:

1. Los costos del año actual (variable ACOVY) para predecir los costos del siguiente año, ya que si un paciente es costoso en el año actual, es de esperarse que también sea costoso en el año siguiente.

2. Los costos obtenidos por el modelo de DxCG (variable COSTDXCG). Este modelo sólo toma en cuenta las variables de condición médica de los pacientes.

La predicción de los *benchmarks* en el 0.5% superior de los datos se muestra en la siguiente tabla:

Año	Resultados en el 0.5% superior	ACOVY	COSTDXCG
1997 29,063 datos en total 145 datos en el 0.5% superior	Número aciertos	33	29
	% Aciertos	22.76%	20%
	Promedio de costos en el siguiente año (millones de dólares)	$12.38	$11.73
1998 38,879 datos en total 194 datos en el 0.5% superior	Número aciertos	37	35
	% Aciertos	19.07%	18.04%
	Promedio de costos en el siguiente año (millones de dólares)	$13.74	$14.35
2000 90,104 datos en total 450 datos en el 0.5% superior	Número aciertos	96	82
	% Aciertos	21.33%	18.22%
	Promedio de costos en el siguiente año (millones de dólares)	$37.36	$35.60

Tabla 4.12 Resultados de la predicción usando las variables ACOVY y COSTDXCG como *benchmarks*.

La tabla muestra los resultados de predecir los pacientes más costosos mediante los dos *benchmarks*. En la columna Año se indican los datos de prueba sobre los que se están prediciendo los pacientes más costosos en el siguiente año. Por ejemplo, la fila 1997 corresponde a la predicción de los pacientes más costosos en 1998 dados los datos de 1997.

Para obtener el desempeño del *benchmark* ACOVY se ordenaron los datos descendentemente por esta variable y se seleccionó el 0.5% superior de los pacientes, es decir el 0.5% con los valores más altos de ACOVY. En este 0.5% de datos se contó cuantos pacientes realmente estaban en el 0.5% de pacientes más costosos del siguiente año. Se siguió el mismo procedimiento para obtener el desempeño del benchmark COSTDXCG.

Los resultados muestran que los dos *benchmarks* tienen desempeños similares, con un desempeño ligeramente mayor de ACOVY sobre COSTDXCG. La razón de que ACOVY sea más predictiva que COSTDXCG puede deberse a que COSTDXCG solamente involucra la condición médica del paciente para predecir sus costos en el siguiente año.

En la sección 4.2.2 se mencionó la creación de una nueva variable calculada mediante la combinación de otras cuatro variables. Los resultados de la predicción del 0.5% superior de los pacientes más costosos del siguiente año usando la variable SCORE se muestran en la siguiente tabla:

Año	Resultados en el 0.5% superior	SCORE
1997 29,063 datos en total 145 datos en el 0.5% superior	Número aciertos	41
	% Aciertos	28.28%
	Promedio de costos en el siguiente año (millones de dólares)	$14.25
1998 38,879 datos en total 194 datos en el 0.5% superior	Número aciertos	43
	% Aciertos	22.16%
	Promedio de costos en el siguiente año (millones de dólares)	$10.26

Año	Resultados en el 0.5% superior	SCORE
2000 90,104 datos en total 450 datos en el 0.5% superior	Número aciertos	118
	% Aciertos	26.22%
	Promedio de costos en el siguiente año (millones de dólares)	$43.05

Tabla 4.13 Resultados de la predicción usando la variable SCORE.

La tabla muestra los resultados de predecir los pacientes más costosos mediante la variable SCORE. En la columna Año se indican los datos de prueba sobre los que se están prediciendo los pacientes más costosos en el siguiente año. Por ejemplo, la fila 1997 corresponde a la predicción de los pacientes más costosos en 1998 dados los datos de 1997.

Es importante hacer notar que se obtuvo un mejor desempeño en la predicción utilizando el enfoque multiperspectiva, ya que mediante la variable SCORE se obtuvieron mejores resultados que con los *benchmarks* ACOVY y COSTDXCG. Mediante la variable SCORE se están combinando las variables que en la perfilación se descubrió que eran más predictivas. Además, también se están integrando distintos modelos de predicción, ya que al utilizar la variable COSTDXCG se está tomando en cuenta un modelo de regresión lineal. En la siguiente sección se mostrarán los resultados obtenidos mediante un algoritmo genético y se compararán con los *benchmarks*.

4.4 Resultados

Antes de indicar los resultados de la predicción del 0.5% superior de pacientes más costosos en el siguiente año, se mostrarán los resultados de diversas corridas del algoritmo genético para determinar los "mejores" parámetros del algoritmo genético.

4.4.1 Parámetros del algoritmo genético

Si bien no existe una manera precisa de determinar cuáles son los parámetros óptimos para un algoritmo genético (se pueden tener muchas combinaciones de valores de parámetros y es prácticamente imposible probar todas las combinaciones), sí se puede seguir una metodología que ayude a establecer valores adecuados de los parámetros del algoritmo genético. La metodología empleada fue la siguiente: para cada parámetro se determinaron *a priori* los valores a probar de los parámetros. Los valores usados para los parámetros fueron los siguientes:

Número de generaciones: 5, 10, 50, 100, 500
Población: 10, 50, 100, 500, 1000
p(mutación): 0.001, 0.01, 0.05, 0.1, 0.2, 0.3
p(cruzamiento): 0, 0.5, 1

Para cada combinación de estos parámetros se realizó una corrida sencilla del algoritmo genético. Para reducir el tiempo de corrida, en el entrenamiento sólo se ocuparon los datos del 5% superior de la variable COSTDXCG. En total fueron 450 corridas, una corrida por combinación de parámetros. Para cada corrida se midió el desempeño de cada uno de los seis criterios del algoritmo genético mencionados en la sección 4.3.4 y se obtuvo el promedio de los desempeños de los seis criterios. Estos desempeños se

midieron en el conjunto de entrenamiento de 1997. La función de aptitud empleada fue $\varepsilon(\boldsymbol{X})$.

Los resultados del desempeño promedio para los diferentes valores de los parámetros se muestran en las siguientes gráficas:

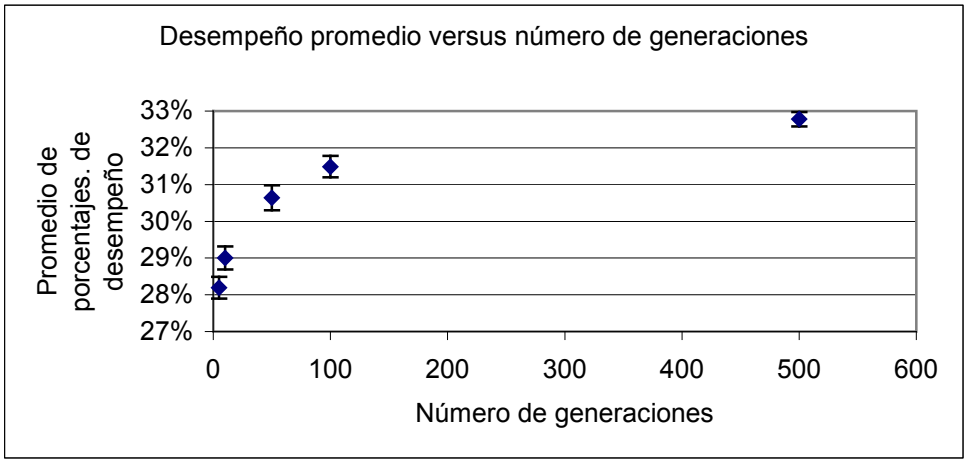

Fig. 4.15 Promedios de porcentajes de desempeño usando diferentes números de generaciones.

En la gráfica del desempeño promedio para distintos números de generaciones se observa que entre más generaciones se tienen el desempeño mejora. Con 50 generaciones se tiene un buen desempeño, por lo que se eligieron 50 generaciones. Con más generaciones se obtienen mejores resultados, sin embargo, los tiempos de corrida del entrenamiento del algoritmo genético también se incrementan, por lo que se optó por usar solamente 50 generaciones.

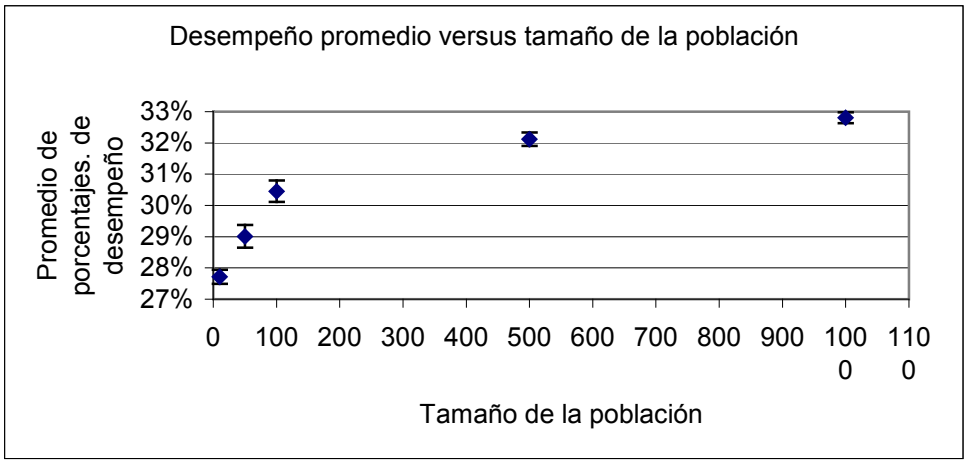

Fig. 4.16 Promedios de porcentajes de desempeño usando diferentes tamaños de poblaciones.

Se observa un comportamiento similar en la gráfica del desempeño promedio para distintos tamaños de poblaciones. Entre más grande sea la población, mejores son los desempeños, con la desventaja del incremento en los tiempos de corridas. En este caso se decidió usar una población de tamaño 100.

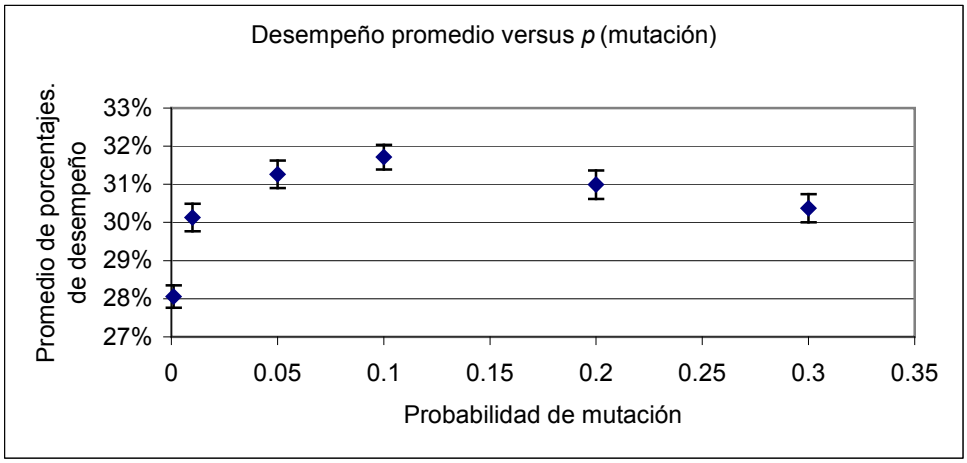

Fig. 4.17 Promedios de porcentajes de desempeño usando diferentes probabilidades de mutación.

El comportamiento observado en la gráfica del desempeño promedio para distintas probabilidades de mutación es diferente. En esa gráfica se observa que se alcanza un desempeño máximo cuando se tiene una probabilidad de mutación de 0.1 y después el desempeño decae. En este caso se optó por usar el mejor valor de *p*(mutación), 0.1.

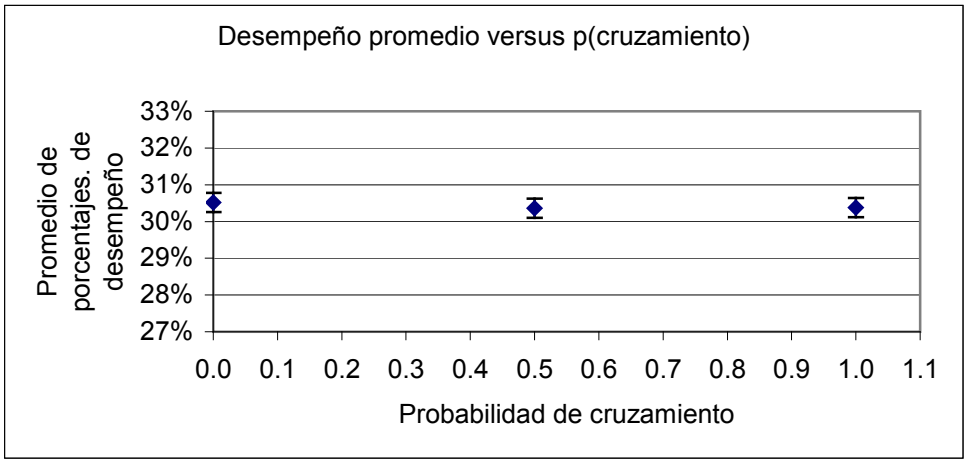

Fig. 4.18 Promedios de los porcentajes de desempeño usando diferentes probabilidades de cruzamiento.

Con la gráfica del desempeño promedio para distintas probabilidades de cruzamiento se tiene otro comportamiento. La gráfica indica que la probabilidad de cruzamiento no afecta notoriamente el desempeño promedio y no se ve una tendencia clara como en los demás parámetros. Dado que no hay una gran diferencia se decidió usar una probabilidad de cruzamiento de 1. Pudiera pensarse que el costo computacional de aplicar probabilidad de cruzamiento 1 es alto, pero no es así. La carga computacional del entrenamiento del algoritmo genético implementado reside en la evaluación de la función de aptitud.

De acuerdo a los resultados anteriores, los parámetros elegidos para correr el algoritmo genético fueron: número generaciones = 50, tamaño población = 100, *p*(mutación) = 0.1 y *p*(cruzamiento) = 1. Se pueden obtener mejores desempeños aumentando el número de

generaciones y el tamaño de la población. Sin embargo, como ya se mencionó, esto a costa de incrementar los tiempos de entrenamiento del algoritmo genético.

4.4.2 Desempeño de los clasificadores

Se hicieron veinte multi-corridas, cada multi-corrida tuvo diez corridas internas. Los parámetros usados fueron: función de aptitud $\varepsilon(\boldsymbol{X})$, número generaciones = 50, tamaño población = 100, p(mutación) = 0.1 y p(cruzamiento) = 1, número de clasificadores = 100. A la lista resultante de clasificadores se le añadieron los clasificadores ACOVY = 0.5% superior, COSTDXCG = 0.5% superior y SCORE = 0.5% superior ya que estos clasificadores mostraron ser altamente predictivos. En el entrenamiento sólo se ocuparon los datos del 5% superior de la variable COSTDXCG para reducir los tiempos de corrida y también para enfocarse en datos con alta probabilidad de pertenecer a la clase, ya que como anteriormente se mencionó, la variable COSTDXCG es una variable predictiva de la clase. Las veinte multi-corridas se realizaron usando como conjuntos de entrenamiento los datos de 1997, 1998 y 2000 de manera independiente. Los resultados se muestran en las siguientes tres tablas. El desempeño se midió usando los distintos criterios del algoritmo genético: AP, aptitud promedio; NC, número de correspondencias; G, ganador; RR, ganador con re-ranqueo; RE, ganador con re-evaluación; RR_RR, ganador con re-ranqueo y re-evaluación.

Promedio del número de aciertos en las veinte multi-corridas						
Datos	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1997	42.25	47.50	47.20	44.75	48.85	49.30
Prueba 1998	44.55	45.45	44.15	45.05	45.05	44.45
Prueba 2000	121.10	113.00	123.00	119.10	120.25	116.90
Promedio del porcentaje de aciertos en las veinte multi-corridas						
Datos	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1997	29.14%	32.76%	32.55%	30.86%	33.69%	34.00%
Prueba 1998	22.96%	23.43%	22.76%	23.22%	23.22%	22.91%
Prueba 2000	26.91%	25.11%	27.33%	26.47%	26.72%	25.98%
Promedio de los costos asociados en las veinte multi-corridas (millones de dólares)						
Datos	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1997	$14.27	$15.26	$ 15.16	$14.53	$15.45	$15.28
Prueba 1998	$15.57	$16.29	$ 15.99	$15.79	$16.15	$15.84
Prueba 2000	$44.02	$42.10	$ 44.56	$43.32	$43.86	$42.83

Tabla 4.14 Resultados de los distintos criterios del algoritmo genético (20 multi-corridas) usando como conjunto de entrenamiento los datos de 1997.

La tabla 4.14 muestra los resultados de predecir los pacientes más costosos mediante los clasificadores encontrados por el algoritmo genético. En la columna Datos se indican los datos sobre los que se están prediciendo los pacientes más costosos en el siguiente año. Por ejemplo, la fila "Entrenamiento 1997" corresponde a la predicción de los pacientes más costosos en 1998 dados los datos de 1997, usando los clasificadores encontrados al entrenar el algoritmo genético con los datos de 1997. Hay que aclarar que los clasificadores se ordenan acuerdo a los datos de entrenamiento y al criterio empleado. Este mismo orden se se usa en los datos de prueba para seleccionar el conjunto de pacientes con mayor probabilidad de estar en el 0.5% superior más costoso. Otro punto a aclarar es que el número de aciertos y los costos asociados mostrados en la tabla 4.14 dependen del número de pacientes en el top 0.5% de los datos: se tienen más pacientes

en el conjunto de datos "Prueba 2000", por eso los promedios de número de aciertos y costos son considerablemente mayores que en los otros dos conjuntos de datos.

Los promedios del porcentaje de aciertos de la tabla 4.14 se muestran en la figura 4.19. En la figura se puede apreciar que el mejor desempeño se obtuvo con el criterio "ganador con re-ranqueo y re-evaluación". Sin embargo, al predecir en los datos de prueba, tanto de 1998, como del 2000, los resultados obtenidos con este criterio no fueron tan buenos como en el conjunto de entrenamiento. Los mejores resultados en los datos de prueba del 2000 se obtuvieron con el criterio "ganador" y en los datos de prueba de 1998 con el criterio "numero de correspondencias".

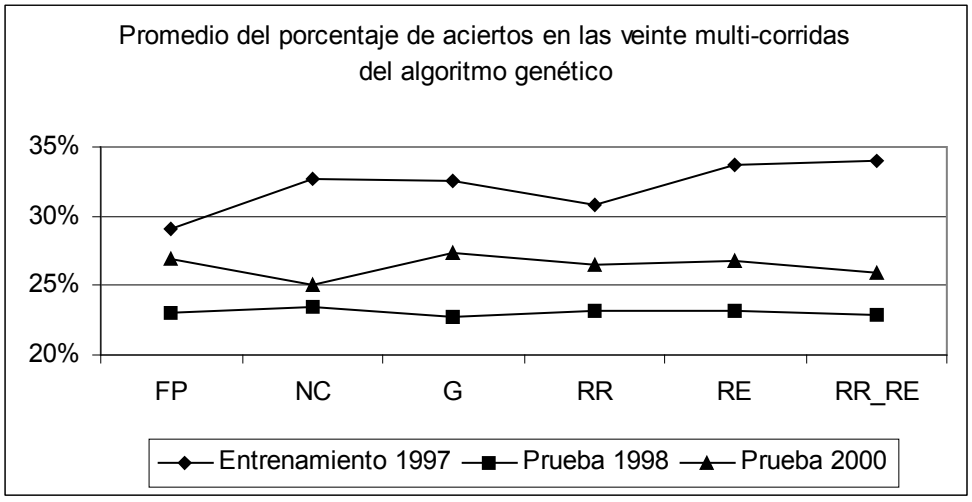

Fig. 4.19 Promedio del porcentaje de aciertos en las veinte multi-corridas del algoritmo genético usando como conjunto de entrenamiento los datos de 1997.

Varianza del número de aciertos en las veinte multi-corridas						
	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1997	6.93	9.32	4.59	11.67	3.40	3.91
Prueba 1998	6.26	7.73	2.24	6.89	1.94	7.31
Prueba 2000	12.20	101.58	7.47	69.36	11.99	74.83
Varianza del porcentaje de aciertos en las veinte multi-corridas						
	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1997	3.30	4.43	2.18	5.55	1.62	1.86
Prueba 1998	1.66	2.05	0.60	1.83	0.52	1.94
Prueba 2000	0.60	5.02	0.37	3.43	0.59	3.70
Varianza del promedio de los costos asociados en las veinte multi-corridas (millones de dólares)						
	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1997	0.28	0.53	0.09	0.71	0.13	0.17
Prueba 1998	0.44	0.21	0.11	0.36	0.06	0.28
Prueba 2000	1.52	6.19	0.89	4.38	1.04	4.58

Tabla 4.15 Varianzas de los resultados de los distintos criterios del algoritmo genético (20 multi-corridas) usando como conjunto de entrenamiento los datos de 1997.

La tabla 4.15 muestra las varianzas de los resultados del número de aciertos, porcentaje de aciertos y promedio de los costos asociados calculados en las veinte multi-corridas del algoritmo genético.

Promedio del número de aciertos en las veinte multi-corridas						
Datos	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1998	42.45	48.20	44.75	47.60	49.20	51.25
Prueba 2000	112.50	112.00	127.55	117.75	94.40	105.25
Promedio del porcentaje de aciertos en las veinte multi-corridas						
Datos	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1998	21.88%	24.85%	23.07%	24.54%	25.36%	26.42%
Prueba 2000	25.00%	24.89%	28.34%	26.17%	20.98%	23.39%
Promedio de los costos asociados en las veinte multi-corridas (millones de dólares)						
Datos	AP	NC	G	RR	RE	RR_RE
Entrenamiento 1998	$15.24	$16.52	$16.14	$16.10	$16.57	$16.45
Prueba 2000	$41.58	$41.60	$45.71	$42.98	$36.73	$39.31

Tabla 4.16 Resultados de los distintos criterios del algoritmo genético usando como conjunto de entrenamiento los datos de 1998.

La tabla 4.16 es muy similar a la tabla 4.15, pero usando los clasificadores encontrados entrenando el algoritmo genético con los datos de 1998.

Los promedios del porcentaje de aciertos de la tabla 4.16 se muestran en la figura 4.20. En la figura se puede apreciar que el mejor criterio en los datos de entrenamiento también fue el criterio "ganador con re-ranqueo y re-evaluación", pero nuevamente, este criterio desmejora en los datos de prueba, tanto de 1998, como del 2000. El mejor criterio en los datos de prueba de 1998 y del 2000 fue el criterio "ganador".

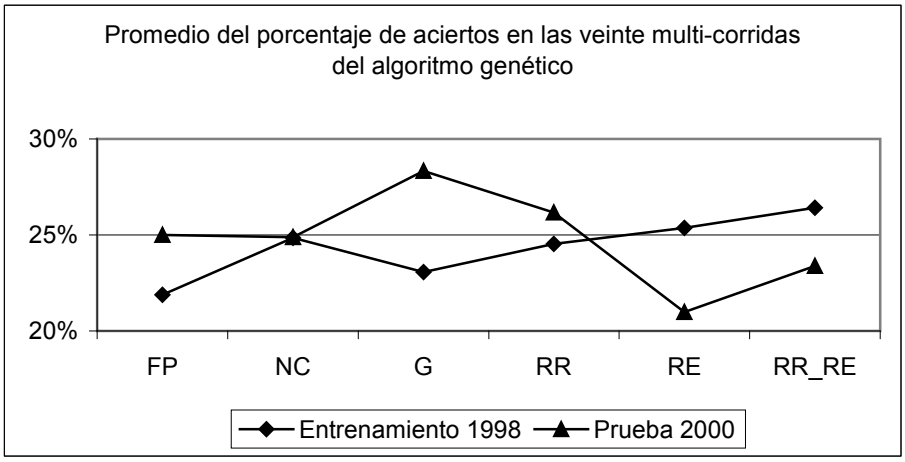

Fig. 4.20 Promedio del porcentaje de aciertos en las veinte multi-corridas del algoritmo genético usando como conjunto de entrenamiento los datos de 1998.

Es de interés notar que los desempeños de la predicción sobre los datos de 1998 siempre fueron los más bajos, independientemente del conjunto de entrenamiento usado. Incluso al usar como conjunto de entrenamiento los datos de 1998, ya que en este caso tres de los seis criterios tuvieron los desempeños más bajos en el conjunto de datos de 1998. La razón de esto podría ser que el desempeño de los clasificadores ACOVY = 0.5% superior,

COSTDXCG = 0.5% superior y SCORE = 0.5% superior fue menor en los datos de 1998, que en los datos de 1997 y del 2000 como se puede apreciar en las tablas 4.12 y 4.13. Dado que estos tres clasificadores siempre se incluyen en la lista de los mejores clasificadores, es natural que tengan gran influencia en la predicción del 0.5% superior y por lo tanto, en el caso de los datos de 1998 hagan decaer el desempeño.

Otro punto de interés fue que en general el criterio "aptitud promedio" tuvo un bajo desempeño con respecto al resto de los criterios. La razón de esto podría ser que al promediar las aptitudes se está restando peso a los mejores clasificadores. En general, el criterio "número de correspondencias" tiene mejor desempeño que el criterio "aptitud promedio". En este criterio también se está diluyendo la contribución de los mejores clasificadores, pero quizás en menor grado dado que no se realizan promedios.

Los criterios "ganador con re-evaluación" y "ganador con re-ranqueo y re-evaluación" tuvieron buenos desempeños en los datos de entrenamiento, sin embargo, no generalizaron bien en los datos de prueba. La razón de esto quizás sea que al hacer la re-evaluación se están sobre-ajustando los resultados, ya que la re-evaluación depende de manera muy estrecha de los datos de entrenamiento. En el caso del criterio "ganador con re-ranqueo" sucede algo similar.

Los resultados obtenidos con el algoritmo genético, entrenado con los datos de 1997, para predecir el 0.5% superior más costoso del 2001 (usando los datos del 2000), fueron mejores que los dos *benchmarks* y en la mayoría de los criterios del algoritmo genético fueron mejores que el resultado obtenido mediante la variable SCORE. Hay que tomar en cuenta que los resultados obtenidos mediante el algoritmo genético fueron usando distintas corridas, en cambio el resultado obtenido mediante la variable SCORE fue solamente mediante una sola clasificación. Además, la ventaja de los algoritmos genéticos es que obtienen un conjunto de clasificadores que proporcionan información que puede servir para tomar decisiones más detalladas que en el caso de la variable SCORE.

En la siguiente tabla se listan algunos de los mejores clasificadores encontrados por el algoritmo genético en las veinte multi-corridas. En la tabla se observa que las variables de costos están muy presentes en los clasificadores. En cuanto a las variables de condición médica se observa que la variable HCC131 (*Renal failure*) también es importante. Otras variables de condición médica presentes en los clasificadores son HCC015 (*Diabetes with Renal Manifestation*), HCC166 (*Major Symptoms, Abnormalities*), HCC036 (*Other Gastrointestinal Disorders*), HCC164 (*Major Complications of Medical Care and Trauma*).

N(C|$\boldsymbol{X}$)	N($\boldsymbol{X}$)	p(C|$\boldsymbol{X}$)	ε	Clasificador
41	120	0.341667	29.439587	COVYO = 6
31	72	0.430556	29.008923	COVYO = 6 COVYO_4 = 6
29	64	0.453125	28.835287	COVYO = 6 HCC131 = 1
40	122	0.327869	28.430107	SCORE = 6
35	100	0.35	27.560118	COVYO = 6 HCC166 = 1
27	61	0.442623	27.476549	COVYO_4 = 6 SCORE = 6
22	41	0.536585	27.476006	COVYO_3 = 6 COVYO_4 = 6
32	85	0.376471	27.416922	SUDDENNESS = 2 COVYO_4 = 6
24	52	0.461538	26.490906	COVYO_4 = 6 HCC131 = 1
26	61	0.42623	26.423203	SCORE = 6 HCC131 = 1

N(C|$\boldsymbol{X}$)	N($\boldsymbol{X}$)	p(C|$\boldsymbol{X}$)	ε	Clasificador
27	66	0.409091	26.339325	COVYO = 6 COVYO_3 = 6
27	67	0.402985	26.126947	SEX = 1 COVYO = 6
33	100	0.33	25.914738	COVYO_4 = 6
23	51	0.45098	25.614643	COVYO = 6 HCC015 = 1
20	39	0.512821	25.576513	ACOVY = 6 COVYO_4 = 6
23	52	0.442308	25.35004	COVYO_3 = 6 SCORE = 6
22	48	0.458333	25.268953	SCORE = 6 HCC015 = 1
16	26	0.615385	25.185612	ACOVY = 6 COVYO_4 = 6 COSTRI4 = 0
25	62	0.403226	25.148756	CONCAVITY = 2 COVYO_4 = 6
32	100	0.32	25.092047	SCORE = 6 HCC166 = 1
20	41	0.487805	24.906353	ACOVY = 6 COVYO_3 = 6
20	41	0.487805	24.906353	COVYO_2 = 6 COVYO_4 = 6
26	69	0.376812	24.725397	COVYO = 6 HOS2 = 1
25	64	0.390625	24.721836	SCORE = 6 HCC036 = 1
24	60	0.4	24.534241	SCORE = 6 HCC183 = 1
19	39	0.487179	24.259155	COVYO_2 = 6 HCC015 = 1
28	83	0.337349	24.160292	COSTRII43 = 0 SCORE = 6
24	62	0.387097	24.103939	OFFSET = 2 SCORE = 6
15	25	0.6	24.063684	COSTRI43 = 0 ACOVY_1 = 6
29	90	0.322222	23.977844	HOS3 = 1 SCORE = 6
28	85	0.329412	23.847591	ACOVY = 6 SCORE = 6
23	59	0.38983	23.686298	ACOVY = 6 COVYO = 6
17	33	0.515152	23.637135	COVYO = 6 COVYO_2 = 6 HCC015 = 1
20	46	0.434783	23.422844	COVYO_3 = 6 COSTRI4 = 0
20	46	0.434783	23.422844	COVYO_2 = 6 COVYO_3 = 6
25	71	0.352113	23.36902	COVYO = 6 HOS1 = 1
13	20	0.65	23.362804	COVYO_4 = 6 ACOVY_1 = 6
17	34	0.5	23.265772	NLR = 1 SCORE = 6
22	56	0.392857	23.262581	SEX = 1 COVYO_4 = 6
21	52	0.403846	23.068308	COVYO = 6 ACOVY_2 = 5
26	80	0.325	22.810928	COVYO_4 = 6 HCC166 = 1
29	99	0.292929	22.750355	COVYO_2 = 6
29	99	0.292929	22.750355	SUDDENNESS = 2 COVYO_2 = 6
16	32	0.5	22.571114	COVYO = 6 HCC179 = 1
19	45	0.422222	22.473669	COVYO_4 = 6 HCC015 = 1
19	45	0.422222	22.473669	COVYO_3 = 6 HCC131 = 1
19	45	0.422222	22.473669	COVYO_2 = 6 COSTRI4 = 0
20	50	0.4	22.396595	COVYO = 6 COSTDXCG = 5
21	55	0.381818	22.380432	HCC015 = 1 HCC131 = 1 HCC164 = 1
25	77	0.324675	22.355698	CONCAVITY = 2 COVYO_2 = 6

Tabla 4.17 50 mejores clasificadores encontrados en veinte multi-corridas del algoritmo genético.

Por último, en la siguiente tabla se muestran los resultados de la clasificación Bayesiana ingenua en el 0.5% superior de pacientes más costosos del siguiente año usando tres

distintos conjuntos de variables: todas las variables, las variables relacionadas con costos y las variables relacionadas con la condición médica (HCC).

	Todas las variables	Variables costos	Variables HCC
Entrenamiento 1997	24.14%	25.52%	20.69%
Prueba 1998	20.62%	18.56%	18.04%
Prueba 2000	20.89%	22.22%	15.11%
	Todas las variables	Variables costos	Variables HCC
Entrenamiento 1998	21.13%	18.04%	19.07%
Prueba 2000	20.89%	21.78%	15.56%

Tabla 4.18 Porcentaje de aciertos en el 0.5% superior mediante la clasificación Bayesiana ingenua.

En la tabla es de interés notar que las variables relacionadas con los costos son más predictivas que las variables relacionadas con la condición médica del paciente. Incluso se obtienen mejores desempeños usando solamente las variables relacionadas con los costos que usando todas las variables. Esto se puede apreciar mejor en la figura 4.22.

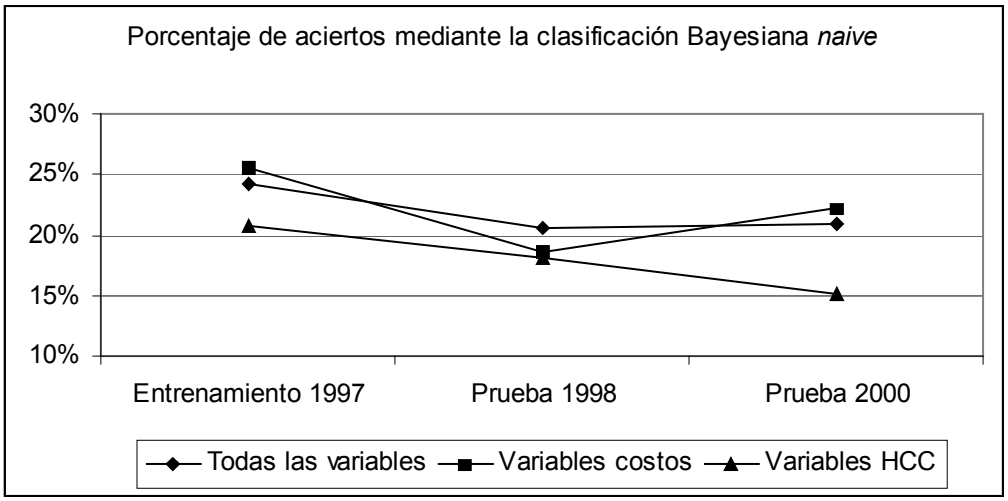

Fig. 4.21 Porcentaje de aciertos mediante la clasificación Bayesiana ingenua.

En el caso de la clasificación Bayesiana ingenua, los resultados fueron mejores que los *benchmarks*, pero no mejores que la variable SCORE, ni que el algoritmo genético. La razón de esto podría ser que la clasificación Bayesiana ingenua es mejor sobre otros modelos cuando se dispone de pocos datos debido a que reduce el problema de la falta de certeza estadística, como se mencionó en la sección 2.5.2.

El enfoque multiperspectiva seguido en este caso de estudio mejoró el desempeño de las predicciones, ya que al incoporar la variable derivada SCORE se superó significativamente el *benchmark* obtenido al usar las variables ACOVY y COSTDXCG para hacer las predicciones. El enfoque multiperspectiva también se empleó para ordenar los clasificadores del algoritmo genético y también se superó el *benchmark* de ACOVY y COSRDXCG. Si bien en algunos casos no se superó el desempeño de la variable SCORE, la ventaja del algoritmo genético es que produce clasificadores que son de fácil interpretación.

Conclusiones

La minería de datos es una disciplina que requiere mucha experiencia para poder ser practicada de manera adecuada. En ocasiones uno quisiera que existiera una receta, si no mágica, sí concreta de cuáles pasos seguir para resolver satisfactoriamente un problema de minería de datos. SIn embargo, esta receta no existe, y menos para los problemas reales de minería de datos.

Resolver un problema real de minería de datos requiere un análisis cuidadoso en todos los pasos del proceso que no es posible automatizar. Este análisis incluye: entender bien las metas y características del problema, involucrarse con el dominio del problema, entender los datos disponibles y sus limitaciones, preparar los datos para facilitar y permitir su análisis, analizar la información que nos dan los datos, así como conocer distintos modelos de minería de datos para elegir aquellos más adecuados al problema. Si bien existen herramientas que pueden facilitar el trabajo, se deben conocer bien sus características para poder emplearlas adecuadamente. Es importante resaltar que estas herramientas no sustituyen el análisis realizado por el minero.

A pesar de que no existe una receta única para resolver problemas de minería de datos, sí hay una serie de aspectos a considerar para resolver este tipo de problemas. Dichos aspectos se presentaron principalmente al final del capítulo uno.

Existe una gran cantidad de métodos de minería de datos. Es imposible abarcarlos todos. Sin embargo, se presentó un panorama general de diversos métodos para las tareas de descripción y predicción que en la práctica han resultado ser muy útiles. Para profundizar más en los métodos presentados se proporcionaron diversas referencias.

La generalización de los problemas de minería de datos como la búsqueda de predictibilidad en un paisaje topográfico es útil porque proporciona un marco teórico que permite comprender diversos aspectos involucrados en la resolución de los problemas de minería de datos, como son: las relaciones más predictivas entre las variables (las que generan picos en el paisaje de predictibilidad), la importancia que tiene la forma de los picos, la selección de modelos de minería de datos dependiendo del paisaje de predictibilidad, etc. En la práctica, en un paisaje de predictibilidad sólo se pueden visualizar relaciones entre muy pocas variables debido a las limitaciones visuales del ser humano. Sin embargo, no por esto el concepto deja de ser útil para explicar y entender varios aspectos de los problemas de minería de datos.

Al adoptar un enfoque multiperspectiva frecuentemente se obtienen mejores resultados de predicción ya que se integran las opiniones de diversos agentes, que pueden ir desde la selección de variables, hasta la combinación de diversos modelos de minería de datos. El enfoque multiperspectiva permite reducir las desventajas de un método en particular así como realzar sus ventajas. En los dos problemas presentados, CoIL y DxCG se obtuvieron mejores resultados de predicción al seguir un enfoque multiperspectiva. Esto se vio con los datos de CoIL al introducir dos nuevas variables que combinaban, cada una, dos variables. Ocurrió lo mismo con los datos de DxCG, en donde se introdujo la variable COSTDXCG, la cual contiene los resultados de un modelo de regresión lineal y también se creó una nueva variable, SCORE, como combinación de cuatro variables existentes, mejorando el desempeño de la predicción de manera significativa. El desempeño del algoritmo genético, el cual a su vez incorporó las variables COSTDXCG y

SCORE, también fue superior a los desempeños de predicción de los *benchmarks* dados por las variables ACOVY y COSTDXCG para predecir el top 0.5% superior de los pacientes más costosos.

Además de mejorar los desempeños de predicción, el enfoque multiperspectiva también se puede emplear para resolver distintos objetivos de un problema. Por ejemplo, en el caso de predicción en los datos de CoIL, los resultados de la perfilación de las variables sociodemográficas se pueden emplear para identificar las áreas geográficas donde hay nuevos clientes con potencial para adquirir pólizas de casas rodantes, mientras que los resultados de predicción son útiles para identificar a clientes existentes pero que todavía no adquieren pólizas de casas rodantes.

Los casos de estudio fueron principalmente problemas de descripción y predicción (realizada mediante clasificación), por lo que el enfoque multiperspectiva se enfocó en este tipo específico de problemas, pero esto no significa que dicho enfoque sólo se pueda llevar a cabo en este tipo de problemas. Así como una persona toma en cuenta diversos factores para resolver un problema complejo, para resolver cualquier problema real de minería de datos se puede adoptar este enfoque multiperspectiva, ya sea un problema de análisis de series de tiempos, análisis de texto, agrupamiento, etc.

Los enfoques multiperspectiva seguidos en los dos casos de estudio no son únicos y muy probablemente haya otras implementaciones mejores. Estas implementaciones dependen del problema a resolver: lo que es bueno para la solución de un problema no necesariamente es bueno para otro problema. Lo que se desea resaltar es el hecho de que el empleo de un enfoque multiperspectiva proporciona diversas ventajas.

Con respecto al trabajo a futuro se pueden hacer mejoras a los diferentes criterios de ordenamiento de los clasificadores encontrados mediante el algoritmo genético, así como tratar de establecer los casos en que funcionan mejor dichos criterios.

Apéndice A: Información sobre datos CoIL

Valor	Descripción
1	Ingresos altos con niños caros
2	Provinciales muy importantes
3	Seniors de alto estatus
4	Departamentos con seniors afluentes
5	Seniors variados
6	Con carrera y niños
7	Doble ingreso y sin niños
8	Familias clase media
9	Familias completas modernas
10	Familias estables
11	Familias que inician
12	Familias jóvenes afluentes
13	Familias jóvenes americanas
14	Jóvenes cosmopolitanos
15	Seniors cosmopolitanos
16	Estudiantes en departamentos
17	Con maestría reciente en la ciudad
18	Juventud soltera
19	Juventud suburbana
20	Diversidad étnica
21	Juventud urbana sin dinero ni posesiones
22	Habitantes variados en departamentos
23	Jóvenes crecientes
24	Jóvenes con bajo nivel educativo
25	Seniors jóvenes en la ciudad
26	Adultos mayores con casa propia
27	Seniors en departamentos
28	Adultos mayores residenciales
29	Seniors sin patio delantero
30	Solteros mayores religiosos
31	Católicos de bajo ingresos
32	Seniors variados
33	Familias grandes de clase baja
34	Familias grandes con niños que trabajan
35	Familias de pueblo
36	Casados con niños
37	Habitantes variados en ciudades pequeñas
38	Familias tradicionales
39	Familias grandes religiosas
40	Granjas de familias grandes
41	Rurales variados

Tabla A.1 Descripciones de valores tipo L0.

Valor	Descripción
1	20 - 30 años
2	30 - 40 años
3	40 - 50 años
4	50 - 60 años
5	60 - 70 años
6	70 - 80 años

Tabla A.2 Descripciones de valores tipo L1.

Valor	Descripción
1	Hedonistas exitosos
2	Cultivadores
3	Familia promedio
4	Solteros con carrera
5	Buen nivel de vida
6	Seniors viajeros
7	Retirados y religiosos
8	Familia con adultos
9	Familias conservadoras
10	Granjeros

Tabla A.3 Descripciones de valores tipo L2.

Valor	Descripción
0	0%
1	1 - 10%
2	11 - 23%
3	24 - 36%
4	37 - 49%
5	50 - 62%
6	63 - 75%
7	76 - 88%
8	89 - 99%
9	100%

Tabla A.4 Descripciones de valores tipo L3.

Valor	Descripción
0	0 f
1	1 – 49 f
2	50 – 99 f
3	100 – 199 f
4	200 – 499 f
5	500 – 999 f
6	1000 – 4999 f
7	5000 – 9999 f
8	10.000 - 19.999 f
9	> 20.000

Tabla A.5 Descripciones de valores tipo L4.

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
PPERSAUT	6	Contribución de pólizas de auto = f 1,000 – 4,999	10.8080	262	2319	348	11.30	75.29
PBRAND	4	Contribución de pólizas de incendio = f 200 – 499	9.3628	151	1226	348	12.32	43.39
APLEZIER	1	Número de pólizas de bote = 1	7.6876	12	31	348	38.71	3.45
MKOOPKLA	7	Clase con poder adquisitivo = 76 - 88%	7.4918	67	474	348	14.14	19.25
MOSTYPE	8	Subtipo de cliente = Familias clase media	7.0419	51	339	348	15.04	14.66
MOSHOOFD	2	Tipo principal de cliente = Cultivadores	6.7765	66	502	348	13.15	18.97
APERSAUT	2	Número de pólizas de auto = 2	6.2653	38	246	348	15.45	10.92
APERSAUT	1	Número de pólizas de auto = 1	6.0665	237	2712	348	8.74	68.10
PWAPART	2	Contribución de seguros contra daño a terceros = f 50 – 99	5.8342	191	2128	348	8.98	54.89
AWAPART	1	Número de seguros contra daño a terceros = 1	5.3688	201	2334	348	8.61	57.76
ABYSTAND	1	Número de pólizas de seguridad social = 1	5.2298	16	81	348	19.75	4.60
MHHUUR	0	Casa rentada = 0%	5.1041	94	949	348	9.91	27.01
MHKOOP	9	Casa propia = 100%	5.1041	94	949	348	9.91	27.01
PPLEZIER	1	Contribución de pólizas de bote = f 1 – 49	5.0956	3	5	348	60.00	0.86
PMOTSCO	3	Contribución de pólizas de motocicleta = f 100 – 199	4.4341	2	3	348	66.67	0.57
PPLEZIER	6	Contribución de pólizas de bote = f 1,000 – 4,999	4.4341	2	3	348	66.67	0.57
MOPLLAAG	2	Educación de nivel bajo = 11 - 23%	4.1047	65	667	348	9.75	18.68
MOPLHOOG	4	Educación de nivel alto = 37 - 49%	4.0917	37	326	348	11.35	10.63
MINKGEM	5	Ingreso promedio = 50 - 62%	4.0799	70	733	348	9.55	20.11
ABRAND	1	Número de pólizas de incendio = 1	3.9676	232	3017	348	7.69	66.67
MSKA	7	Clase social A = 76 - 88%	3.9286	13	79	348	16.46	3.74
MBERARBG	1	Trabajadores capacitados = 1 - 10%	3.8848	83	921	348	9.01	23.85
MAUT1	7	1 auto = 76 - 88%	3.8760	119	1413	348	8.42	34.20
MAUT0	0	Sin auto = 0%	3.8028	121	1450	348	8.34	34.77
MGODPR	7	Protestante = 76 - 88%	3.7811	55	564	348	9.75	15.80
PGEZONG	3	Contribución de pólizas de seguro contra accidentes familiares = f 100 – 199	3.7706	4	13	348	30.77	1.15
PPLEZIER	4	Contribución de pólizas de bote = f 200 – 499	3.7706	4	13	348	30.77	1.15
ALEVEN	4	Número de seguros de vida = 4	3.7610	3	8	348	37.50	0.86
MOSTYPE	12	Subtipo de cliente = Familias jóvenes afluentes	3.7496	16	111	348	14.41	4.60
MBERHOOG	7	Estatus alto = 76 - 88%	3.7385	14	92	348	15.22	4.02
MBERMIDD	7	Administración media = 76 - 88%	3.5918	22	178	348	12.36	6.32
MSKC	1	Clase social C = 1 - 10%	3.5147	30	272	348	11.03	8.62
MBERHOOG	6	Estatus alto = 63 - 75%	3.5015	18	138	348	13.04	5.17
MSKA	4	Clase social A = 37 - 49%	3.4985	29	261	348	11.11	8.33

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
MINK4575	4	Ingresos 45 - 75,000 = 37 - 49%	3.4362	88	1034	348	8.51	25.29
PBYSTAND	4	Contribución de pólizas de seguridad social = 4	3.4149	8	44	348	18.18	2.30
PBYSTAND	2	Contribución de pólizas de seguridad social = 2	3.3800	4	15	348	26.67	1.15
MINKGEM	7	Ingreso promedio = 76 - 88%	3.3795	17	131	348	12.98	4.89
MOPLLAAG	1	Educación de nivel bajo = 1 - 10%	3.3758	27	243	348	11.11	7.76
MINKM30	2	Ingresos < 30,000 = 11 - 23%	3.2659	91	1094	348	8.32	26.15
PPLEZIER	2	Contribución de pólizas de bote = f 50 – 99	3.2091	2	5	348	40.00	0.57
PPLEZIER	3	Contribución de pólizas de bote = f 100 – 199	3.2091	2	5	348	40.00	0.57
MOPLHOOG	5	Educación de nivel alto = 50 - 62%	3.0299	21	187	348	11.23	6.03
MSKB1	6	Clase social B1 = 63 - 75%	2.9575	5	25	348	20.00	1.44
MOPLMIDD	7	Educación de nivel medio = 76 - 88%	2.9005	18	157	348	11.47	5.17
MSKB1	3	Clase social B1 = 24 - 36%	2.8298	65	775	348	8.39	18.68
MRELOV	0	Otra relación = 0%	2.8187	93	1173	348	7.93	26.72
MRELGE	9	Casado = 100%	2.7754	66	794	348	8.31	18.97
PWAOREG	6	Contribución de pólizas de seguro contra discapacidades = f 1,000 – 4,999	2.7719	4	19	348	21.05	1.15
AWAOREG	1	Número de pólizas de seguro contra discapacidades = 1	2.7719	4	19	348	21.05	1.15
MINKGEM	4	Ingreso promedio = 37 - 49%	2.7607	139	1854	348	7.50	39.94
MINKM30	1	Ingresos < 30,000 = 1 - 10%	2.7465	54	630	348	8.57	15.52
MBERARBO	0	Trabajadores no capacitados = 0%	2.7305	78	968	348	8.06	22.41
MSKC	2	Clase social C = 11 - 23%	2.7168	71	870	348	8.16	20.40
MOPLLAAG	0	Educación de nivel bajo = 0%	2.7146	29	299	348	9.70	8.33
MOSTYPE	3	Subtipo de cliente = Seniors de alto estatus	2.7043	25	249	348	10.04	7.18
MSKA	6	Clase social A = 63 - 75%	2.6958	12	96	348	12.50	3.45
MOSHOOFD	1	Tipo principal de cliente = Hedonistas exitosos	2.6940	48	552	348	8.70	13.79
MSKD	0	Clase social D = 0%	2.6578	188	2607	348	7.21	54.02
PZEILPL	1	Contribución de pólizas de tablas de surfeo = f 1 – 49	2.6262	1	2	348	50.00	0.29
APLEZIER	2	Número de pólizas de bote = 2	2.6262	1	2	348	50.00	0.29
AFIETS	3	Número de pólizas de bicicleta = 3	2.6262	1	2	348	50.00	0.29
AGEZONG	1	Número de pólizas de seguro contra accidentes familiares = 1	2.5514	6	38	348	15.79	1.72
PBYSTAND	3	Contribución de pólizas de seguridad social = 3	2.4147	4	22	348	18.18	1.15
MSKA	5	Clase social A = 50 - 62%	2.3989	14	127	348	11.02	4.02
MBERARBO	1	Trabajadores no capacitados = 1 - 10%	2.3476	76	980	348	7.76	21.84
MINKM30	0	Ingresos < 30,000 = 0%	2.3428	98	1304	348	7.52	28.16
PLEVEN	4	Contribución de seguros de vida = f 200 – 499	2.3413	11	94	348	11.70	3.16

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(C)	P(C|***X***)	P(***X***|C)
MHHUUR	1	Casa rentada = 1 - 10%	2.3279	37	428	348	8.64	10.63
MHKOOP	8	Casa propia = 89 - 99%	2.3279	37	428	348	8.64	10.63
MFALLEEN	0	Soltero = 0%	2.3124	128	1757	348	7.29	36.78
MBERBOER	0	Granjero = 0%	2.2446	284	4176	348	6.80	81.61
MINK7512	2	Ingresos 75 - 122,000 = 11 - 23%	2.1779	58	736	348	7.88	16.67
PFIETS	1	Contribución de pólizas de bicicleta = f 1 – 49	2.1617	15	147	348	10.20	4.31
MINK4575	5	Ingresos 45 - 75,000 = 50 - 62%	2.1233	41	498	348	8.23	11.78
MOSTYPE	1	Subtipo de cliente = Ingresos altos con niños caros	2.1168	13	124	348	10.48	3.74
MZFONDS	0	Servicio nacional de salud = 0%	2.1116	7	55	348	12.73	2.01
MZPART	9	Segura privado de salud = 100%	2.1116	7	55	348	12.73	2.01
MZFONDS	2	Servicio nacional de salud = 11 - 23%	2.0824	27	307	348	8.79	7.76
MZPART	7	Segura privado de salud = 76 - 88%	2.0824	27	307	348	8.79	7.76
MOPLHOOG	6	Educación de nivel alto = 63 - 75%	2.0589	8	67	348	11.94	2.30
MGODGE	1	Sin religión = 1 - 10%	2.0171	21	230	348	9.13	6.03
AZEILPL	1	Número de pólizas de tablas de surfeo = 1	1.9987	1	3	348	33.33	0.29
MBERHOOG	4	Estatus alto = 37 - 49%	1.9625	33	397	348	8.31	9.48
MKOOPKLA	8	Clase con poder adquisitivo = 89 - 99%	1.9490	35	426	348	8.22	10.06
MGODGE	0	Sin religión = 0%	1.9247	37	456	348	8.11	10.63
MINKGEM	8	Ingreso promedio = 89 - 99%	1.9239	8	70	348	11.43	2.30
PINBOED	1	Contribución de pólizas de seguro de propiedad = f 1 – 49	1.9130	3	18	348	16.67	0.86
MOSTYPE	6	Subtipo de cliente = Con carrera y niños	1.8897	12	119	348	10.08	3.45
MSKC	8	Clase social C = 89 - 99%	1.8803	8	71	348	11.27	2.30
PAANHANG	2	Contribución de pólizas de tráiler = f 50 – 99	1.8672	5	38	348	13.16	1.44
MZFONDS	5	Servicio nacional de salud = 50 - 62%	1.8626	72	974	348	7.39	20.69
MFGEKIND	6	Familia con hijos = 63 - 75%	1.8394	27	321	348	8.41	7.76
MGODOV	2	Otra religión = 11 - 23%	1.8155	99	1388	348	7.13	28.45
MINK7512	4	Ingresos 75 - 122,000 = 37 - 49%	1.8138	14	147	348	9.52	4.02
PBRAND	3	Contribución de pólizas de incendio = f 100 – 199	1.8091	68	920	348	7.39	19.54
MINK7512	1	Ingresos 75 - 122,000 = 1 - 10%	1.8043	97	1359	348	7.14	27.87
MOPLHOOG	7	Educación de nivel alto = 76 - 88%	1.7434	6	51	348	11.76	1.72
PWAPART	3	Contribución de seguros contra daño a terceros = f 100 – 199	1.7074	2	11	348	18.18	0.57
ALEVEN	3	Número de seguros de vida = 3	1.7074	2	11	348	18.18	0.57
MKOOPKLA	6	Clase con poder adquisitivo = 63 - 75%	1.7066	66	901	348	7.33	18.97
MZPART	4	Segura privado de salud = 37 -	1.7016	72	992	348	7.26	20.69

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
		49%						
ALEVEN	2	Número de seguros de vida = 2	1.6969	10	100	348	10.00	2.87
MFWEKIND	8	Familia sin hijos = 89 - 99%	1.6713	18	206	348	8.74	5.17
MAUT1	9	1 auto = 100%	1.6546	39	505	348	7.72	11.21
MINK3045	9	Ingresos 30 - 45,000 = 100%	1.6098	9	90	348	10.00	2.59
MINK7512	9	Ingresos 75 - 122,000 = 100%	1.6048	1	4	348	25.00	0.29
MGODRK	2	Católico romano = 11 - 23%	1.5870	54	733	348	7.37	15.52
MRELGE	7	Casado = 76 - 88%	1.5836	116	1683	348	6.89	33.33
PMOTSCO	5	Contribución de pólizas de motocicleta = f 500 – 999	1.5564	4	32	348	12.50	1.15
MOPLMIDD	5	Educación de nivel medio = 50 - 62%	1.5353	54	738	348	7.32	15.52
MBERMIDD	9	Administración media = 100%	1.5177	8	80	348	10.00	2.30
AINBOED	1	Número de pólizas de seguro de propiedad = 1	1.5071	5	44	348	11.36	1.44
MZFONDS	4	Servicio nacional de salud = 37 - 49%	1.4871	28	357	348	7.84	8.05
MBERMIDD	5	Administración media = 50 - 62%	1.4706	33	431	348	7.66	9.48
PWABEDR	3	Contribución de seguros contra daño a terceros (empresas) = f 100 – 199	1.4295	3	23	348	13.04	0.86
AFIETS	2	Número de pólizas de bicicleta = 2	1.4235	4	34	348	11.76	1.15
MZFONDS	3	Servicio nacional de salud = 24 - 36%	1.4015	15	177	348	8.47	4.31
MFWEKIND	6	Familia sin hijos = 63 - 75%	1.3865	56	783	348	7.15	16.09
MZPART	5	Segura privado de salud = 50 - 62%	1.3802	28	364	348	7.69	8.05
MINK7512	5	Ingresos 75 - 122,000 = 50 - 62%	1.3797	7	71	348	9.86	2.01
MZPART	6	Segura privado de salud = 63 - 75%	1.3786	15	178	348	8.43	4.31
MINK4575	3	Ingresos 45 - 75,000 = 24 - 36%	1.3766	84	1215	348	6.91	24.14
MGEMOMV	4	Habitantes promedio por casa = 4	1.3744	50	693	348	7.22	14.37
PLEVEN	5	Contribución de seguros de vida = f 500 – 999	1.3604	4	35	348	11.43	1.15
AAANHANG	1	Número de pólizas de tráiler = 1	1.3583	6	59	348	10.17	1.72
MBERARBG	8	Trabajadores capacitados = 89 - 99%	1.3479	3	24	348	12.50	0.86
AFIETS	1	Número de pólizas de bicicleta = 1	1.3473	10	111	348	9.01	2.87
MINK4575	6	Ingresos 45 - 75,000 = 63 - 75%	1.3312	11	125	348	8.80	3.16
MSKB2	4	Clase social B2 = 37 - 49%	1.3262	47	652	348	7.21	13.51
MBERMIDD	6	Administración media = 63 - 75%	1.2742	17	211	348	8.06	4.89
MRELSA	0	Unión libre = 0%	1.2511	161	2448	348	6.58	46.26
MGODGE	7	Sin religión = 76 - 88%	1.2436	9	101	348	8.91	2.59
MBERZELF	1	Empresario = 1 - 10%	1.2352	82	1202	348	6.82	23.56
MGODRK	1	Católico romano = 1 - 10%	1.2049	107	1599	348	6.69	30.75

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
MBERHOOG	3	Estatus alto = 24 - 36%	1.1984	53	756	348	7.01	15.23
MBERHOOG	8	Estatus alto = 89 - 99%	1.1961	3	26	348	11.54	0.86
MBERZELF	2	Empresario = 11 - 23%	1.1756	26	348	348	7.47	7.47
MINK7512	3	Ingresos 75 - 122,000 = 24 - 36%	1.1553	19	246	348	7.72	5.46
MAUT0	2	Sin auto = 11 - 23%	1.1373	108	1625	348	6.65	31.03
MSKA	3	Clase social A = 24 - 36%	1.1371	48	685	348	7.01	13.79
MOPLHOOG	3	Educación de nivel alto = 24 - 36%	1.1370	39	547	348	7.13	11.21
PINBOED	2	Contribución de pólizas de seguro de propiedad = f 50 – 99	1.1006	2	16	348	12.50	0.57
MGEMOMV	3	Habitantes promedio por casa = 3	1.0529	171	2646	348	6.46	49.14
MOPLLAAG	3	Educación de nivel bajo = 24 - 36%	1.0279	47	680	348	6.91	13.51
MFALLEEN	1	Soltero = 1 - 10%	0.9788	64	951	348	6.73	18.39
MFGEKIND	7	Familia con hijos = 76 - 88%	0.9737	8	96	348	8.33	2.30
MINK123M	2	Ingresos > 123,000 = 11 - 23%	0.9737	8	96	348	8.33	2.30
MGODOV	0	Otra religión = 0%	0.9684	130	2003	348	6.49	37.36
MOPLMIDD	9	Educación de nivel medio = 100%	0.9197	4	43	348	9.30	1.15
PBROM	0	Contribución de pólizas de ciclomotor = f 0	0.8974	340	5426	348	6.27	97.70
ABROM	0	Número de pólizas de ciclomotor = 0	0.8974	340	5426	348	6.27	97.70
MBERARBG	0	Trabajadores capacitados = 0%	0.8946	77	1167	348	6.60	22.13
MFWEKIND	9	Familia sin hijos = 100%	0.8914	14	186	348	7.53	4.02
MAUT1	8	1 auto = 89 - 99%	0.8875	19	261	348	7.28	5.46
MOSHOOFD	3	Tipo principal de cliente = Familia promedio	0.8561	59	886	348	6.66	16.95
MGODPR	5	Protestante = 50 - 62%	0.7927	97	1501	348	6.46	27.87
MSKB1	8	Clase social B1 = 89 - 99%	0.7782	1	8	348	12.50	0.29
MOSTYPE	37	Subtipo de cliente = Habitantes variados en ciudades pequeñas	0.7747	10	132	348	7.58	2.87
MGODGE	2	Sin religión = 11 - 23%	0.7713	69	1055	348	6.54	19.83
MRELGE	8	Casado = 89 - 99%	0.7597	25	361	348	6.93	7.18
MAUT2	2	2 autos = 11 - 23%	0.7583	112	1748	348	6.41	32.18
MOSTYPE	13	Subtipo de cliente = Familias jóvenes americanas	0.7253	13	179	348	7.26	3.74
MOSTYPE	36	Subtipo de cliente = Casados con niños	0.7174	16	225	348	7.11	4.60
MSKC	0	Clase social C = 0%	0.7169	25	364	348	6.87	7.18
MGODOV	3	Otra religión = 24 - 36%	0.6942	18	257	348	7.00	5.17
AMOTSCO	1	Número de pólizas de motocicleta = 1	0.6934	15	211	348	7.11	4.31
MINK3045	3	Ingresos 30 - 45,000 = 24 - 36%	0.6776	74	1147	348	6.45	21.26
MINK123M	1	Ingresos > 123,000 = 1 - 10%	0.6709	50	763	348	6.55	14.37
MOPLHOOG	9	Educación de nivel alto = 100%	0.6497	1	9	348	11.11	0.29
MBERARBO	8	Trabajadores no capacitados = 89 - 99%	0.6497	1	9	348	11.11	0.29
MAUT2	6	2 autos = 63 - 75%	0.6497	1	9	348	11.11	0.29

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
PTRACTOR	6	Contribución de pólizas de tractor = f 1,000 – 4,999	0.6497	1	9	348	11.11	0.29
MRELOV	3	Otra relación = 24 - 36%	0.6390	74	1152	348	6.42	21.26
MINK4575	7	Ingresos 45 - 75,000 = 76 - 88%	0.6303	7	93	348	7.53	2.01
MOSTYPE	38	Subtipo de cliente = Familias tradicionales	0.6270	23	339	348	6.78	6.61
MINKGEM	6	Ingreso promedio = 63 - 75%	0.6225	24	355	348	6.76	6.90
MOPLHOOG	8	Educación de nivel alto = 89 - 99%	0.6160	2	22	348	9.09	0.57
MHHUUR	3	Casa rentada = 24 - 36%	0.6157	39	593	348	6.58	11.21
MBERMIDD	4	Administración media = 37 - 49%	0.5515	61	953	348	6.40	17.53
MOPLMIDD	8	Educación de nivel medio = 89 - 99%	0.5467	3	37	348	8.11	0.86
MBERZELF	3	Empresario = 24 - 36%	0.5467	3	37	348	8.11	0.86
MBERARBG	9	Trabajadores capacitados = 100%	0.5366	1	10	348	10.00	0.29
AMOTSCO	2	Número de pólizas de motocicleta = 2	0.5366	1	10	348	10.00	0.29
MAANTHUI	2	Número de casas = 2	0.5283	33	505	348	6.53	9.48
MOSTYPE	2	Subtipo de cliente = Provinciales muy importantes	0.5118	6	82	348	7.32	1.72
MINK3045	1	Ingresos 30 - 45,000 = 1 - 10%	0.5104	18	268	348	6.72	5.17
MRELOV	2	Otra relación = 11 - 23%	0.5072	110	1756	348	6.26	31.61
MHKOOP	6	Casa propia = 63 - 75%	0.4972	39	604	348	6.46	11.21
MSKD	1	Clase social D = 1 - 10%	0.4881	98	1563	348	6.27	28.16
MBERARBG	7	Trabajadores capacitados = 76 - 88%	0.4785	5	68	348	7.35	1.44
MFWEKIND	7	Familia sin hijos = 76 - 88%	0.4547	23	351	348	6.55	6.61
PLEVEN	3	Contribución de seguros de vida = f 100 – 199	0.4506	6	84	348	7.14	1.72
MOSTYPE	20	Subtipo de cliente = Diversidad étnica	0.4266	2	25	348	8.00	0.57
PGEZONG	2	Contribución de pólizas de seguro contra accidentes familiares = f 50 – 99	0.4266	2	25	348	8.00	0.57
MFGEKIND	1	Familia con hijos = 1 - 10%	0.3859	24	372	348	6.45	6.90
MBERARBG	2	Trabajadores capacitados = 11 - 23%	0.3850	86	1382	348	6.22	24.71
MOSHOOFD	9	Tipo principal de cliente = Familias conservadoras	0.3481	42	667	348	6.30	12.07
MBERZELF	4	Empresario = 37 - 49%	0.3443	1	12	348	8.33	0.29
APERSAUT	3	Número de pólizas de auto = 3	0.3443	1	12	348	8.33	0.29
MSKB2	3	Clase social B2 = 24 - 36%	0.3404	73	1175	348	6.21	20.98
MINK3045	4	Ingresos 30 - 45,000 = 37 - 49%	0.3376	84	1356	348	6.19	24.14
PMOTSCO	4	Contribución de pólizas de motocicleta = f 200 – 499	0.3150	9	136	348	6.62	2.59
MSKB1	4	Clase social B1 = 37 - 49%	0.2902	19	298	348	6.38	5.46
MGEMLEEF	3	Edad promedio = 40 - 50 años	0.2834	183	3000	348	6.10	52.59
MGODRK	6	Católico romano = 63 - 75%	0.2608	1	13	348	7.69	0.29
MSKA	8	Clase social A = 89 - 99%	0.2608	1	13	348	7.69	0.29
MSKD	7	Clase social D = 76 - 88%	0.2608	1	13	348	7.69	0.29

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
PTRACTOR	5	Contribución de pólizas de tractor = f 500 – 999	0.2602	2	28	348	7.14	0.57
MAUT2	4	2 autos = 37 - 49%	0.2451	19	301	348	6.31	5.46
MAUT0	1	Sin auto = 1 - 10%	0.2447	48	776	348	6.19	13.79
MOSTYPE	7	Subtipo de cliente = Doble ingreso y sin niños	0.2353	3	44	348	6.82	0.86
PWALAND	0	Contribución de seguros contra daño a terceros (agricultura) = f 0	0.2331	345	5702	348	6.05	99.14
AWALAND	0	Número de seguros contra daño a terceros (agricultura) = 0	0.2331	345	5702	348	6.05	99.14
MOPLHOOG	2	Educación de nivel alto = 11 - 23%	0.2020	70	1144	348	6.12	20.11
PTRACTOR	0	Contribución de pólizas de tractor = f 0	0.1986	343	5679	348	6.04	98.56
ATRACTOR	0	Número de pólizas de tractor = 0	0.1986	343	5679	348	6.04	98.56
MFGEKIND	8	Familia con hijos = 89 - 99%	0.1840	1	14	348	7.14	0.29
MFGEKIND	0	Familia con hijos = 0%	0.1805	23	371	348	6.20	6.61
MFWEKIND	2	Familia sin hijos = 11 - 23%	0.1748	39	635	348	6.14	11.21
PWABEDR	2	Contribución de seguros contra daño a terceros (empresas) = f 50 – 99	0.1593	2	30	348	6.67	0.57
MINK3045	6	Ingresos 30 - 45,000 = 63 - 75%	0.1533	25	406	348	6.16	7.18
MRELOV	1	Otra relación = 1 - 10%	0.1421	33	539	348	6.12	9.48
MGEMLEEF	5	Edad promedio = 60 - 70 años	0.1408	12	193	348	6.22	3.45
MRELGE	6	Casado = 63 - 75%	0.1165	71	1172	348	6.06	20.40
MFGEKIND	4	Familia con hijos = 37 - 49%	0.1139	88	1455	348	6.05	25.29
MFWEKIND	3	Familia sin hijos = 24 - 36%	0.1137	59	973	348	6.06	16.95
MSKB2	6	Clase social B2 = 63 - 75%	0.1127	6	96	348	6.25	1.72
MBERHOOG	9	Estatus alto = 100%	0.1114	2	31	348	6.45	0.57
MOPLMIDD	4	Educación de nivel medio = 37 - 49%	0.0853	86	1426	348	6.03	24.71
AWABEDR	1	Número de seguros contra daño a terceros (empresas) = 1	0.0742	5	81	348	6.17	1.44
MGODPR	2	Protestante = 11 - 23%	0.0699	24	396	348	6.06	6.90
PWERKT	0	Contribución de pólizas de máquinas para agricultura = f 0	0.0695	348	5801	348	6.00	100.00
AWERKT	0	Número de pólizas de máquinas para agricultura = 0	0.0695	348	5801	348	6.00	100.00
MOPLMIDD	3	Educación de nivel medio = 24 - 36%	0.0580	80	1330	348	6.02	22.99
MFALLEEN	2	Soltero = 11 - 23%	0.0553	75	1247	348	6.01	21.55
PBESAUT	0	Contribución de pólizas de camioneta repartidora = f 0	0.0482	346	5774	348	5.99	99.43
ABESAUT	0	Número de pólizas de camioneta repartidora = 0	0.0482	346	5774	348	5.99	99.43
PPERSONG	0	Contribución de pólizas de seguro contra accidentes = f 0	0.0473	347	5791	348	5.99	99.71
APERSONG	0	Número de pólizas de seguro contra accidentes = 0	0.0473	347	5791	348	5.99	99.71
PBROM	5	Contribución de pólizas de ciclomotor = f 500 – 999	0.0460	1	16	348	6.25	0.29

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
MHHUUR	6	Casa rentada = 63 - 75%	0.0360	23	382	348	6.02	6.61
MHKOOP	3	Casa propia = 24 - 36%	0.0360	23	382	348	6.02	6.61
PVRAAUT	0	Contribución de pólizas de camión de carga = f 0	0.0298	348	5813	348	5.99	100.00
AVRAAUT	0	Número de pólizas de camión de carga = 0	0.0298	348	5813	348	5.99	100.00
MGEMLEEF	2	Edad promedio = 30 - 40 años	0.0232	87	1452	348	5.99	25.00
MOPLLAAG	4	Educación de nivel bajo = 37 - 49%	0.0192	51	851	348	5.99	14.66
MGODGE	3	Sin religión = 24 - 36%	0.0165	87	1453	348	5.99	25.00
MAANTHUI	1	Número de casas = 1	0.0101	315	5267	348	5.98	90.52
MINK3045	2	Ingresos 30 - 45,000 = 11 - 23%	0.0095	55	919	348	5.98	15.80
MOSTYPE	14	Subtipo de cliente = Jóvenes cosmopolitanos	0.0000	0	0	348	-100.00	0.00
MAANTHUI	9	Número de casas = 9	0.0000	0	0	348	-100.00	0.00
MGEMOMV	6	Habitantes promedio por casa = 6	0.0000	0	0	348	-100.00	0.00
MGODOV	6	Otra religión = 63 - 75%	0.0000	0	0	348	-100.00	0.00
MGODOV	7	Otra religión = 76 - 88%	0.0000	0	0	348	-100.00	0.00
MGODOV	8	Otra religión = 89 - 99%	0.0000	0	0	348	-100.00	0.00
MGODOV	9	Otra religión = 100%	0.0000	0	0	348	-100.00	0.00
MRELSA	8	Unión libre = 89 - 99%	0.0000	0	0	348	-100.00	0.00
MRELSA	9	Unión libre = 100%	0.0000	0	0	348	-100.00	0.00
MBERZELF	6	Empresario = 63 - 75%	0.0000	0	0	348	-100.00	0.00
MBERZELF	7	Empresario = 76 - 88%	0.0000	0	0	348	-100.00	0.00
MBERZELF	8	Empresario = 89 - 99%	0.0000	0	0	348	-100.00	0.00
MBERZELF	9	Empresario = 100%	0.0000	0	0	348	-100.00	0.00
MSKD	8	Clase social D = 89 - 99%	0.0000	0	0	348	-100.00	0.00
MAUT2	8	2 autos = 89 - 99%	0.0000	0	0	348	-100.00	0.00
MAUT2	9	2 autos = 100%	0.0000	0	0	348	-100.00	0.00
MINK123M	6	Ingresos > 123,000 = 63 - 75%	0.0000	0	0	348	-100.00	0.00
MINK123M	8	Ingresos > 123,000 = 89 - 99%	0.0000	0	0	348	-100.00	0.00
MKOOPKLA	0	Clase con poder adquisitivo = 0%	0.0000	0	0	348	-100.00	0.00
MKOOPKLA	9	Clase con poder adquisitivo = 100%	0.0000	0	0	348	-100.00	0.00
PWAPART	4	Contribución de seguros contra daño a terceros = f 200 – 499	0.0000	0	0	348	-100.00	0.00
PWAPART	5	Contribución de seguros contra daño a terceros = f 500 – 999	0.0000	0	0	348	-100.00	0.00
PWAPART	6	Contribución de seguros contra daño a terceros = f 1,000 – 4,999	0.0000	0	0	348	-100.00	0.00
PWAPART	7	Contribución de seguros contra daño a terceros = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PWAPART	8	Contribución de seguros contra daño a terceros = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PWAPART	9	Contribución de seguros contra daño a terceros = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PWABEDR	7	Contribución de seguros contra	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
		daño a terceros (empresas) = f 5,000 – 9,999						
PWABEDR	8	Contribución de seguros contra daño a terceros (empresas) = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PWABEDR	9	Contribución de seguros contra daño a terceros (empresas) = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PWALAND	1	Contribución de seguros contra daño a terceros (agricultura) = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PWALAND	5	Contribución de seguros contra daño a terceros (agricultura) = f 500 – 999	0.0000	0	0	348	-100.00	0.00
PWALAND	6	Contribución de seguros contra daño a terceros (agricultura) = f 1,000 – 4,999	0.0000	0	0	348	-100.00	0.00
PWALAND	7	Contribución de seguros contra daño a terceros (agricultura) = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PWALAND	8	Contribución de seguros contra daño a terceros (agricultura) = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PWALAND	9	Contribución de seguros contra daño a terceros (agricultura) = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PPERSAUT	1	Contribución de pólizas de auto = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PPERSAUT	2	Contribución de pólizas de auto = f 50 – 99	0.0000	0	0	348	-100.00	0.00
PPERSAUT	3	Contribución de pólizas de auto = f 100 – 199	0.0000	0	0	348	-100.00	0.00
PPERSAUT	9	Contribución de pólizas de auto = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PBESAUT	1	Contribución de pólizas de camioneta repartidora = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PBESAUT	2	Contribución de pólizas de camioneta repartidora = f 50 – 99	0.0000	0	0	348	-100.00	0.00
PBESAUT	3	Contribución de pólizas de camioneta repartidora = f 100 – 199	0.0000	0	0	348	-100.00	0.00
PBESAUT	4	Contribución de pólizas de camioneta repartidora = f 200 – 499	0.0000	0	0	348	-100.00	0.00
PBESAUT	8	Contribución de pólizas de camioneta repartidora = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PBESAUT	9	Contribución de pólizas de camioneta repartidora = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PMOTSCO	1	Contribución de pólizas de motocicleta = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PMOTSCO	2	Contribución de pólizas de motocicleta = f 50 – 99	0.0000	0	0	348	-100.00	0.00
PMOTSCO	8	Contribución de pólizas de motocicleta = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
PMOTSCO	9	Contribución de pólizas de motocicleta = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PVRAAUT	1	Contribución de pólizas de camión de carga = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PVRAAUT	2	Contribución de pólizas de camión de carga = f 50 – 99	0.0000	0	0	348	-100.00	0.00
PVRAAUT	3	Contribución de pólizas de camión de carga = f 100 – 199	0.0000	0	0	348	-100.00	0.00
PVRAAUT	5	Contribución de pólizas de camión de carga = f 500 – 999	0.0000	0	0	348	-100.00	0.00
PVRAAUT	7	Contribución de pólizas de camión de carga = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PVRAAUT	8	Contribución de pólizas de camión de carga = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PAANHANG	6	Contribución de pólizas de tráiler = f 1,000 – 4,999	0.0000	0	0	348	-100.00	0.00
PAANHANG	7	Contribución de pólizas de tráiler = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PAANHANG	8	Contribución de pólizas de tráiler = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PAANHANG	9	Contribución de pólizas de tráiler = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PTRACTOR	1	Contribución de pólizas de tractor = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PTRACTOR	2	Contribución de pólizas de tractor = f 50 – 99	0.0000	0	0	348	-100.00	0.00
PTRACTOR	7	Contribución de pólizas de tractor = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PTRACTOR	8	Contribución de pólizas de tractor = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PTRACTOR	9	Contribución de pólizas de tractor = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PWERKT	1	Contribución de pólizas de máquinas para agricultura = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PWERKT	5	Contribución de pólizas de máquinas para agricultura = f 500 – 999	0.0000	0	0	348	-100.00	0.00
PWERKT	7	Contribución de pólizas de máquinas para agricultura = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PWERKT	8	Contribución de pólizas de máquinas para agricultura = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PWERKT	9	Contribución de pólizas de máquinas para agricultura = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PBROM	1	Contribución de pólizas de ciclomotor = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PBROM	7	Contribución de pólizas de ciclomotor = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PBROM	8	Contribución de pólizas de ciclomotor = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PBROM	9	Contribución de pólizas de ciclomotor = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PPERSONG	7	Contribución de pólizas de	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
		seguro contra accidentes = f 5,000 – 9,999						
PPERSONG	8	Contribución de pólizas de seguro contra accidentes = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PPERSONG	9	Contribución de pólizas de seguro contra accidentes = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PGEZONG	1	Contribución de pólizas de seguro contra accidentes familiares = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PGEZONG	4	Contribución de pólizas de seguro contra accidentes familiares = f 200 – 499	0.0000	0	0	348	-100.00	0.00
PGEZONG	5	Contribución de pólizas de seguro contra accidentes familiares = f 500 – 999	0.0000	0	0	348	-100.00	0.00
PGEZONG	6	Contribución de pólizas de seguro contra accidentes familiares = f 1,000 – 4,999	0.0000	0	0	348	-100.00	0.00
PGEZONG	7	Contribución de pólizas de seguro contra accidentes familiares = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PGEZONG	8	Contribución de pólizas de seguro contra accidentes familiares = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PGEZONG	9	Contribución de pólizas de seguro contra accidentes familiares = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PWAOREG	1	Contribución de pólizas de seguro contra discapacidades = f 1 – 49	0.0000	0	0	348	-100.00	0.00
PWAOREG	2	Contribución de pólizas de seguro contra discapacidades = f 50 – 99	0.0000	0	0	348	-100.00	0.00
PWAOREG	3	Contribución de pólizas de seguro contra discapacidades = f 100 – 199	0.0000	0	0	348	-100.00	0.00
PWAOREG	8	Contribución de pólizas de seguro contra discapacidades = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PWAOREG	9	Contribución de pólizas de seguro contra discapacidades = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PBRAND	9	Contribución de pólizas de incendio = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PZEILPL	2	Contribución de pólizas de tablas de surfeo = f 50 – 99	0.0000	0	0	348	-100.00	0.00
PZEILPL	4	Contribución de pólizas de tablas de surfeo = f 200 – 499	0.0000	0	0	348	-100.00	0.00
PZEILPL	5	Contribución de pólizas de tablas de surfeo = f 500 – 999	0.0000	0	0	348	-100.00	0.00
PZEILPL	6	Contribución de pólizas de tablas de surfeo = f 1,000 – 4,999	0.0000	0	0	348	-100.00	0.00
PZEILPL	7	Contribución de pólizas de tablas de surfeo = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
PZEILPL	8	Contribución de pólizas de tablas de surfeo = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PZEILPL	9	Contribución de pólizas de tablas de surfeo = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PPLEZIER	7	Contribución de pólizas de bote = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PPLEZIER	8	Contribución de pólizas de bote = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PPLEZIER	9	Contribución de pólizas de bote = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PFIETS	2	Contribución de pólizas de bicicleta = f 50 – 99	0.0000	0	0	348	-100.00	0.00
PFIETS	3	Contribución de pólizas de bicicleta = f 100 – 199	0.0000	0	0	348	-100.00	0.00
PFIETS	4	Contribución de pólizas de bicicleta = f 200 – 499	0.0000	0	0	348	-100.00	0.00
PFIETS	5	Contribución de pólizas de bicicleta = f 500 – 999	0.0000	0	0	348	-100.00	0.00
PFIETS	6	Contribución de pólizas de bicicleta = f 1,000 – 4,999	0.0000	0	0	348	-100.00	0.00
PFIETS	7	Contribución de pólizas de bicicleta = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PFIETS	8	Contribución de pólizas de bicicleta = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PFIETS	9	Contribución de pólizas de bicicleta = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PINBOED	7	Contribución de pólizas de seguro de propiedad = f 5,000 – 9,999	0.0000	0	0	348	-100.00	0.00
PINBOED	8	Contribución de pólizas de seguro de propiedad = f 10,000 - 19,999	0.0000	0	0	348	-100.00	0.00
PINBOED	9	Contribución de pólizas de seguro de propiedad = f 20,000 - ?	0.0000	0	0	348	-100.00	0.00
PBYSTAND	1	Contribución de pólizas de seguridad social = 1	0.0000	0	0	348	-100.00	0.00
PBYSTAND	6	Contribución de pólizas de seguridad social = 6	0.0000	0	0	348	-100.00	0.00
PBYSTAND	7	Contribución de pólizas de seguridad social = 7	0.0000	0	0	348	-100.00	0.00
PBYSTAND	8	Contribución de pólizas de seguridad social = 8	0.0000	0	0	348	-100.00	0.00
PBYSTAND	9	Contribución de pólizas de seguridad social = 9	0.0000	0	0	348	-100.00	0.00
AWAPART	3	Número de seguros contra daño a terceros = 3	0.0000	0	0	348	-100.00	0.00
AWAPART	4	Número de seguros contra daño a terceros = 4	0.0000	0	0	348	-100.00	0.00
AWAPART	5	Número de seguros contra daño a terceros = 5	0.0000	0	0	348	-100.00	0.00
AWAPART	6	Número de seguros contra daño a terceros = 6	0.0000	0	0	348	-100.00	0.00
AWAPART	7	Número de seguros contra daño a terceros = 7	0.0000	0	0	348	-100.00	0.00
AWAPART	8	Número de seguros contra daño a terceros = 8	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
AWAPART	9	Número de seguros contra daño a terceros = 9	0.0000	0	0	348	-100.00	0.00
AWAPART	10	Número de seguros contra daño a terceros = 10	0.0000	0	0	348	-100.00	0.00
AWAPART	11	Número de seguros contra daño a terceros = 11	0.0000	0	0	348	-100.00	0.00
AWAPART	12	Número de seguros contra daño a terceros = 12	0.0000	0	0	348	-100.00	0.00
AWABEDR	2	Número de seguros contra daño a terceros (empresas) = 2	0.0000	0	0	348	-100.00	0.00
AWABEDR	3	Número de seguros contra daño a terceros (empresas) = 3	0.0000	0	0	348	-100.00	0.00
AWABEDR	4	Número de seguros contra daño a terceros (empresas) = 4	0.0000	0	0	348	-100.00	0.00
AWABEDR	6	Número de seguros contra daño a terceros (empresas) = 6	0.0000	0	0	348	-100.00	0.00
AWABEDR	7	Número de seguros contra daño a terceros (empresas) = 7	0.0000	0	0	348	-100.00	0.00
AWABEDR	8	Número de seguros contra daño a terceros (empresas) = 8	0.0000	0	0	348	-100.00	0.00
AWABEDR	9	Número de seguros contra daño a terceros (empresas) = 9	0.0000	0	0	348	-100.00	0.00
AWABEDR	10	Número de seguros contra daño a terceros (empresas) = 10	0.0000	0	0	348	-100.00	0.00
AWABEDR	11	Número de seguros contra daño a terceros (empresas) = 11	0.0000	0	0	348	-100.00	0.00
AWABEDR	12	Número de seguros contra daño a terceros (empresas) = 12	0.0000	0	0	348	-100.00	0.00
AWALAND	2	Número de seguros contra daño a terceros (agricultura) = 2	0.0000	0	0	348	-100.00	0.00
AWALAND	3	Número de seguros contra daño a terceros (agricultura) = 3	0.0000	0	0	348	-100.00	0.00
AWALAND	4	Número de seguros contra daño a terceros (agricultura) = 4	0.0000	0	0	348	-100.00	0.00
AWALAND	5	Número de seguros contra daño a terceros (agricultura) = 5	0.0000	0	0	348	-100.00	0.00
AWALAND	6	Número de seguros contra daño a terceros (agricultura) = 6	0.0000	0	0	348	-100.00	0.00
AWALAND	7	Número de seguros contra daño a terceros (agricultura) = 7	0.0000	0	0	348	-100.00	0.00
AWALAND	8	Número de seguros contra daño a terceros (agricultura) = 8	0.0000	0	0	348	-100.00	0.00
AWALAND	9	Número de seguros contra daño a terceros (agricultura) = 9	0.0000	0	0	348	-100.00	0.00
AWALAND	10	Número de seguros contra daño a terceros (agricultura) = 10	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
AWALAND	11	Número de seguros contra daño a terceros (agricultura) = 11	0.0000	0	0	348	-100.00	0.00
AWALAND	12	Número de seguros contra daño a terceros (agricultura) = 12	0.0000	0	0	348	-100.00	0.00
APERSAUT	5	Número de pólizas de auto = 5	0.0000	0	0	348	-100.00	0.00
APERSAUT	8	Número de pólizas de auto = 8	0.0000	0	0	348	-100.00	0.00
APERSAUT	9	Número de pólizas de auto = 9	0.0000	0	0	348	-100.00	0.00
APERSAUT	10	Número de pólizas de auto = 10	0.0000	0	0	348	-100.00	0.00
APERSAUT	11	Número de pólizas de auto = 11	0.0000	0	0	348	-100.00	0.00
APERSAUT	12	Número de pólizas de auto = 12	0.0000	0	0	348	-100.00	0.00
ABESAUT	5	Número de pólizas de camioneta repartidora = 5	0.0000	0	0	348	-100.00	0.00
ABESAUT	6	Número de pólizas de camioneta repartidora = 6	0.0000	0	0	348	-100.00	0.00
ABESAUT	7	Número de pólizas de camioneta repartidora = 7	0.0000	0	0	348	-100.00	0.00
ABESAUT	8	Número de pólizas de camioneta repartidora = 8	0.0000	0	0	348	-100.00	0.00
ABESAUT	9	Número de pólizas de camioneta repartidora = 9	0.0000	0	0	348	-100.00	0.00
ABESAUT	10	Número de pólizas de camioneta repartidora = 10	0.0000	0	0	348	-100.00	0.00
ABESAUT	11	Número de pólizas de camioneta repartidora = 11	0.0000	0	0	348	-100.00	0.00
ABESAUT	12	Número de pólizas de camioneta repartidora = 12	0.0000	0	0	348	-100.00	0.00
AMOTSCO	3	Número de pólizas de motocicleta = 3	0.0000	0	0	348	-100.00	0.00
AMOTSCO	4	Número de pólizas de motocicleta = 4	0.0000	0	0	348	-100.00	0.00
AMOTSCO	5	Número de pólizas de motocicleta = 5	0.0000	0	0	348	-100.00	0.00
AMOTSCO	6	Número de pólizas de motocicleta = 6	0.0000	0	0	348	-100.00	0.00
AMOTSCO	7	Número de pólizas de motocicleta = 7	0.0000	0	0	348	-100.00	0.00
AMOTSCO	9	Número de pólizas de motocicleta = 9	0.0000	0	0	348	-100.00	0.00
AMOTSCO	10	Número de pólizas de motocicleta = 10	0.0000	0	0	348	-100.00	0.00
AMOTSCO	11	Número de pólizas de motocicleta = 11	0.0000	0	0	348	-100.00	0.00
AMOTSCO	12	Número de pólizas de motocicleta = 12	0.0000	0	0	348	-100.00	0.00
AVRAAUT	4	Número de pólizas de camión de carga = 4	0.0000	0	0	348	-100.00	0.00
AVRAAUT	5	Número de pólizas de camión de carga = 5	0.0000	0	0	348	-100.00	0.00
AVRAAUT	6	Número de pólizas de camión de carga = 6	0.0000	0	0	348	-100.00	0.00
AVRAAUT	7	Número de pólizas de camión de carga = 7	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
AVRAAUT	8	Número de pólizas de camión de carga = 8	0.0000	0	0	348	-100.00	0.00
AVRAAUT	9	Número de pólizas de camión de carga = 9	0.0000	0	0	348	-100.00	0.00
AVRAAUT	10	Número de pólizas de camión de carga = 10	0.0000	0	0	348	-100.00	0.00
AVRAAUT	11	Número de pólizas de camión de carga = 11	0.0000	0	0	348	-100.00	0.00
AVRAAUT	12	Número de pólizas de camión de carga = 12	0.0000	0	0	348	-100.00	0.00
AAANHANG	4	Número de pólizas de tráiler = 4	0.0000	0	0	348	-100.00	0.00
AAANHANG	5	Número de pólizas de tráiler = 5	0.0000	0	0	348	-100.00	0.00
AAANHANG	6	Número de pólizas de tráiler = 6	0.0000	0	0	348	-100.00	0.00
AAANHANG	7	Número de pólizas de tráiler = 7	0.0000	0	0	348	-100.00	0.00
AAANHANG	8	Número de pólizas de tráiler = 8	0.0000	0	0	348	-100.00	0.00
AAANHANG	9	Número de pólizas de tráiler = 9	0.0000	0	0	348	-100.00	0.00
AAANHANG	10	Número de pólizas de tráiler = 10	0.0000	0	0	348	-100.00	0.00
AAANHANG	11	Número de pólizas de tráiler = 11	0.0000	0	0	348	-100.00	0.00
AAANHANG	12	Número de pólizas de tráiler = 12	0.0000	0	0	348	-100.00	0.00
ATRACTOR	5	Número de pólizas de tractor = 5	0.0000	0	0	348	-100.00	0.00
ATRACTOR	6	Número de pólizas de tractor = 6	0.0000	0	0	348	-100.00	0.00
ATRACTOR	7	Número de pólizas de tractor = 7	0.0000	0	0	348	-100.00	0.00
ATRACTOR	8	Número de pólizas de tractor = 8	0.0000	0	0	348	-100.00	0.00
ATRACTOR	9	Número de pólizas de tractor = 9	0.0000	0	0	348	-100.00	0.00
ATRACTOR	10	Número de pólizas de tractor = 10	0.0000	0	0	348	-100.00	0.00
ATRACTOR	11	Número de pólizas de tractor = 11	0.0000	0	0	348	-100.00	0.00
ATRACTOR	12	Número de pólizas de tractor = 12	0.0000	0	0	348	-100.00	0.00
AWERKT	4	Número de pólizas de máquinas para agricultura = 4	0.0000	0	0	348	-100.00	0.00
AWERKT	5	Número de pólizas de máquinas para agricultura = 5	0.0000	0	0	348	-100.00	0.00
AWERKT	7	Número de pólizas de máquinas para agricultura = 7	0.0000	0	0	348	-100.00	0.00
AWERKT	8	Número de pólizas de máquinas para agricultura = 8	0.0000	0	0	348	-100.00	0.00
AWERKT	9	Número de pólizas de máquinas para agricultura = 9	0.0000	0	0	348	-100.00	0.00
AWERKT	10	Número de pólizas de máquinas para agricultura = 10	0.0000	0	0	348	-100.00	0.00
AWERKT	11	Número de pólizas de máquinas para agricultura = 11	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
AWERKT	12	Número de pólizas de máquinas para agricultura = 12	0.0000	0	0	348	-100.00	0.00
ABROM	3	Número de pólizas de ciclomotor = 3	0.0000	0	0	348	-100.00	0.00
ABROM	4	Número de pólizas de ciclomotor = 4	0.0000	0	0	348	-100.00	0.00
ABROM	5	Número de pólizas de ciclomotor = 5	0.0000	0	0	348	-100.00	0.00
ABROM	6	Número de pólizas de ciclomotor = 6	0.0000	0	0	348	-100.00	0.00
ABROM	7	Número de pólizas de ciclomotor = 7	0.0000	0	0	348	-100.00	0.00
ABROM	8	Número de pólizas de ciclomotor = 8	0.0000	0	0	348	-100.00	0.00
ABROM	9	Número de pólizas de ciclomotor = 9	0.0000	0	0	348	-100.00	0.00
ABROM	10	Número de pólizas de ciclomotor = 10	0.0000	0	0	348	-100.00	0.00
ABROM	11	Número de pólizas de ciclomotor = 11	0.0000	0	0	348	-100.00	0.00
ABROM	12	Número de pólizas de ciclomotor = 12	0.0000	0	0	348	-100.00	0.00
ALEVEN	5	Número de seguros de vida = 5	0.0000	0	0	348	-100.00	0.00
ALEVEN	6	Número de seguros de vida = 6	0.0000	0	0	348	-100.00	0.00
ALEVEN	7	Número de seguros de vida = 7	0.0000	0	0	348	-100.00	0.00
ALEVEN	9	Número de seguros de vida = 9	0.0000	0	0	348	-100.00	0.00
ALEVEN	10	Número de seguros de vida = 10	0.0000	0	0	348	-100.00	0.00
ALEVEN	11	Número de seguros de vida = 11	0.0000	0	0	348	-100.00	0.00
ALEVEN	12	Número de seguros de vida = 12	0.0000	0	0	348	-100.00	0.00
APERSONG	2	Número de pólizas de seguro contra accidentes = 2	0.0000	0	0	348	-100.00	0.00
APERSONG	3	Número de pólizas de seguro contra accidentes = 3	0.0000	0	0	348	-100.00	0.00
APERSONG	4	Número de pólizas de seguro contra accidentes = 4	0.0000	0	0	348	-100.00	0.00
APERSONG	5	Número de pólizas de seguro contra accidentes = 5	0.0000	0	0	348	-100.00	0.00
APERSONG	6	Número de pólizas de seguro contra accidentes = 6	0.0000	0	0	348	-100.00	0.00
APERSONG	7	Número de pólizas de seguro contra accidentes = 7	0.0000	0	0	348	-100.00	0.00
APERSONG	8	Número de pólizas de seguro contra accidentes = 8	0.0000	0	0	348	-100.00	0.00
APERSONG	9	Número de pólizas de seguro contra accidentes = 9	0.0000	0	0	348	-100.00	0.00
APERSONG	10	Número de pólizas de seguro contra accidentes = 10	0.0000	0	0	348	-100.00	0.00
APERSONG	11	Número de pólizas de seguro contra accidentes = 11	0.0000	0	0	348	-100.00	0.00
APERSONG	12	Número de pólizas de seguro contra accidentes = 12	0.0000	0	0	348	-100.00	0.00
AGEZONG	2	Número de pólizas de seguro contra accidentes familiares = 2	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
AGEZONG	3	Número de pólizas de seguro contra accidentes familiares = 3	0.0000	0	0	348	-100.00	0.00
AGEZONG	4	Número de pólizas de seguro contra accidentes familiares = 4	0.0000	0	0	348	-100.00	0.00
AGEZONG	5	Número de pólizas de seguro contra accidentes familiares = 5	0.0000	0	0	348	-100.00	0.00
AGEZONG	6	Número de pólizas de seguro contra accidentes familiares = 6	0.0000	0	0	348	-100.00	0.00
AGEZONG	7	Número de pólizas de seguro contra accidentes familiares = 7	0.0000	0	0	348	-100.00	0.00
AGEZONG	8	Número de pólizas de seguro contra accidentes familiares = 8	0.0000	0	0	348	-100.00	0.00
AGEZONG	9	Número de pólizas de seguro contra accidentes familiares = 9	0.0000	0	0	348	-100.00	0.00
AGEZONG	10	Número de pólizas de seguro contra accidentes familiares = 10	0.0000	0	0	348	-100.00	0.00
AGEZONG	11	Número de pólizas de seguro contra accidentes familiares = 11	0.0000	0	0	348	-100.00	0.00
AGEZONG	12	Número de pólizas de seguro contra accidentes familiares = 12	0.0000	0	0	348	-100.00	0.00
AWAOREG	3	Número de pólizas de seguro contra discapacidades = 3	0.0000	0	0	348	-100.00	0.00
AWAOREG	4	Número de pólizas de seguro contra discapacidades = 4	0.0000	0	0	348	-100.00	0.00
AWAOREG	5	Número de pólizas de seguro contra discapacidades = 5	0.0000	0	0	348	-100.00	0.00
AWAOREG	6	Número de pólizas de seguro contra discapacidades = 6	0.0000	0	0	348	-100.00	0.00
AWAOREG	7	Número de pólizas de seguro contra discapacidades = 7	0.0000	0	0	348	-100.00	0.00
AWAOREG	8	Número de pólizas de seguro contra discapacidades = 8	0.0000	0	0	348	-100.00	0.00
AWAOREG	9	Número de pólizas de seguro contra discapacidades = 9	0.0000	0	0	348	-100.00	0.00
AWAOREG	10	Número de pólizas de seguro contra discapacidades = 10	0.0000	0	0	348	-100.00	0.00
AWAOREG	11	Número de pólizas de seguro contra discapacidades = 11	0.0000	0	0	348	-100.00	0.00
AWAOREG	12	Número de pólizas de seguro contra discapacidades = 12	0.0000	0	0	348	-100.00	0.00
ABRAND	6	Número de pólizas de incendio = 6	0.0000	0	0	348	-100.00	0.00
ABRAND	8	Número de pólizas de incendio = 8	0.0000	0	0	348	-100.00	0.00
ABRAND	9	Número de pólizas de incendio = 9	0.0000	0	0	348	-100.00	0.00
ABRAND	10	Número de pólizas de incendio = 10	0.0000	0	0	348	-100.00	0.00
ABRAND	11	Número de pólizas de incendio = 11	0.0000	0	0	348	-100.00	0.00
ABRAND	12	Número de pólizas de incendio = 12	0.0000	0	0	348	-100.00	0.00
AZEILPL	2	Número de pólizas de tablas de surfeo = 2	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
AZEILPL	3	Número de pólizas de tablas de surfeo = 3	0.0000	0	0	348	-100.00	0.00
AZEILPL	4	Número de pólizas de tablas de surfeo = 4	0.0000	0	0	348	-100.00	0.00
AZEILPL	5	Número de pólizas de tablas de surfeo = 5	0.0000	0	0	348	-100.00	0.00
AZEILPL	6	Número de pólizas de tablas de surfeo = 6	0.0000	0	0	348	-100.00	0.00
AZEILPL	7	Número de pólizas de tablas de surfeo = 7	0.0000	0	0	348	-100.00	0.00
AZEILPL	8	Número de pólizas de tablas de surfeo = 8	0.0000	0	0	348	-100.00	0.00
AZEILPL	9	Número de pólizas de tablas de surfeo = 9	0.0000	0	0	348	-100.00	0.00
AZEILPL	10	Número de pólizas de tablas de surfeo = 10	0.0000	0	0	348	-100.00	0.00
AZEILPL	11	Número de pólizas de tablas de surfeo = 11	0.0000	0	0	348	-100.00	0.00
AZEILPL	12	Número de pólizas de tablas de surfeo = 12	0.0000	0	0	348	-100.00	0.00
APLEZIER	3	Número de pólizas de bote = 3	0.0000	0	0	348	-100.00	0.00
APLEZIER	4	Número de pólizas de bote = 4	0.0000	0	0	348	-100.00	0.00
APLEZIER	5	Número de pólizas de bote = 5	0.0000	0	0	348	-100.00	0.00
APLEZIER	6	Número de pólizas de bote = 6	0.0000	0	0	348	-100.00	0.00
APLEZIER	7	Número de pólizas de bote = 7	0.0000	0	0	348	-100.00	0.00
APLEZIER	8	Número de pólizas de bote = 8	0.0000	0	0	348	-100.00	0.00
APLEZIER	9	Número de pólizas de bote = 9	0.0000	0	0	348	-100.00	0.00
APLEZIER	10	Número de pólizas de bote = 10	0.0000	0	0	348	-100.00	0.00
APLEZIER	11	Número de pólizas de bote = 11	0.0000	0	0	348	-100.00	0.00
APLEZIER	12	Número de pólizas de bote = 12	0.0000	0	0	348	-100.00	0.00
AFIETS	4	Número de pólizas de bicicleta = 4	0.0000	0	0	348	-100.00	0.00
AFIETS	5	Número de pólizas de bicicleta = 5	0.0000	0	0	348	-100.00	0.00
AFIETS	6	Número de pólizas de bicicleta = 6	0.0000	0	0	348	-100.00	0.00
AFIETS	7	Número de pólizas de bicicleta = 7	0.0000	0	0	348	-100.00	0.00
AFIETS	8	Número de pólizas de bicicleta = 8	0.0000	0	0	348	-100.00	0.00
AFIETS	9	Número de pólizas de bicicleta = 9	0.0000	0	0	348	-100.00	0.00
AFIETS	10	Número de pólizas de bicicleta = 10	0.0000	0	0	348	-100.00	0.00
AFIETS	11	Número de pólizas de bicicleta = 11	0.0000	0	0	348	-100.00	0.00
AFIETS	12	Número de pólizas de bicicleta = 12	0.0000	0	0	348	-100.00	0.00
AINBOED	3	Número de pólizas de seguro de propiedad = 3	0.0000	0	0	348	-100.00	0.00
AINBOED	4	Número de pólizas de seguro de propiedad = 4	0.0000	0	0	348	-100.00	0.00
AINBOED	5	Número de pólizas de seguro	0.0000	0	0	348	-100.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
		de propiedad = 5						
AINBOED	6	Número de pólizas de seguro de propiedad = 6	0.0000	0	0	348	-100.00	0.00
AINBOED	7	Número de pólizas de seguro de propiedad = 7	0.0000	0	0	348	-100.00	0.00
AINBOED	8	Número de pólizas de seguro de propiedad = 8	0.0000	0	0	348	-100.00	0.00
AINBOED	9	Número de pólizas de seguro de propiedad = 9	0.0000	0	0	348	-100.00	0.00
AINBOED	10	Número de pólizas de seguro de propiedad = 10	0.0000	0	0	348	-100.00	0.00
AINBOED	11	Número de pólizas de seguro de propiedad = 11	0.0000	0	0	348	-100.00	0.00
AINBOED	12	Número de pólizas de seguro de propiedad = 12	0.0000	0	0	348	-100.00	0.00
ABYSTAND	3	Número de pólizas de seguridad social = 3	0.0000	0	0	348	-100.00	0.00
ABYSTAND	4	Número de pólizas de seguridad social = 4	0.0000	0	0	348	-100.00	0.00
ABYSTAND	5	Número de pólizas de seguridad social = 5	0.0000	0	0	348	-100.00	0.00
ABYSTAND	6	Número de pólizas de seguridad social = 6	0.0000	0	0	348	-100.00	0.00
ABYSTAND	7	Número de pólizas de seguridad social = 7	0.0000	0	0	348	-100.00	0.00
ABYSTAND	8	Número de pólizas de seguridad social = 8	0.0000	0	0	348	-100.00	0.00
ABYSTAND	9	Número de pólizas de seguridad social = 9	0.0000	0	0	348	-100.00	0.00
ABYSTAND	10	Número de pólizas de seguridad social = 10	0.0000	0	0	348	-100.00	0.00
ABYSTAND	11	Número de pólizas de seguridad social = 11	0.0000	0	0	348	-100.00	0.00
ABYSTAND	12	Número de pólizas de seguridad social = 12	0.0000	0	0	348	-100.00	0.00
PWABEDR	0	Contribución de seguros contra daño a terceros (empresas) = f 0	-0.0055	343	5740	348	5.98	98.56
AWABEDR	0	Número de seguros contra daño a terceros (empresas) = 0	-0.0055	343	5740	348	5.98	98.56
MGEMLEEF	4	Edad promedio = 50 - 60 años	-0.0176	64	1073	348	5.96	18.39
MSKB2	2	Clase social B2 = 11 - 23%	-0.0185	100	1676	348	5.97	28.74
MSKC	6	Clase social C = 63 - 75%	-0.0209	29	487	348	5.95	8.33
MGEMOMV	5	Habitantes promedio por casa = 5	-0.0330	4	68	348	5.88	1.15
PZEILPL	0	Contribución de pólizas de tablas de surfeo = f 0	-0.0454	347	5819	348	5.96	99.71
AZEILPL	0	Número de pólizas de tablas de surfeo = 0	-0.0454	347	5819	348	5.96	99.71
MOSTYPE	11	Subtipo de cliente = Familias que inician	-0.0496	9	153	348	5.88	2.59
MFGEKIND	3	Familia con hijos = 24 - 36%	-0.0589	89	1498	348	5.94	25.57
MBERZELF	5	Empresario = 50 - 62%	-0.0633	3	52	348	5.77	0.86
MINK3045	8	Ingresos 30 - 45,000 = 89 - 99%	-0.0656	2	35	348	5.71	0.57
PBESAUT	6	Contribución de pólizas de	-0.0656	2	35	348	5.71	0.57

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
		camioneta repartidora = f 1,000 – 4,999						
MGODRK	5	Católico romano = 50 - 62%	-0.0755	1	18	348	5.56	0.29
PPERSONG	2	Contribución de pólizas de seguro contra accidentes = f 50 – 99	-0.0755	1	18	348	5.56	0.29
MINK4575	8	Ingresos 45 - 75,000 = 89 - 99%	-0.0973	3	53	348	5.66	0.86
PAANHANG	0	Contribución de pólizas de tráiler = f 0	-0.1176	342	5757	348	5.94	98.28
AAANHANG	0	Número de pólizas de tráiler = 0	-0.1176	342	5757	348	5.94	98.28
MRELSA	1	Unión libre = 1 - 10%	-0.1254	120	2030	348	5.91	34.48
PINBOED	0	Contribución de pólizas de seguro de propiedad = f 0	-0.1282	343	5777	348	5.94	98.56
AINBOED	0	Número de pólizas de seguro de propiedad = 0	-0.1282	343	5777	348	5.94	98.56
PAANHANG	1	Contribución de pólizas de tráiler = f 1 – 49	-0.1313	1	19	348	5.26	0.29
MOSTYPE	39	Subtipo de cliente = Familias grandes religiosas	-0.1411	19	328	348	5.79	5.46
MAUT0	6	Sin auto = 63 - 75%	-0.1430	5	89	348	5.62	1.44
PWAOREG	0	Contribución de pólizas de seguro contra discapacidades = f 0	-0.1454	344	5799	348	5.93	98.85
AWAOREG	0	Número de pólizas de seguro contra discapacidades = 0	-0.1454	344	5799	348	5.93	98.85
MOSTYPE	32	Subtipo de cliente = Seniors variados	-0.1521	8	141	348	5.67	2.30
PMOTSCO	0	Contribución de pólizas de motocicleta = f 0	-0.1539	332	5600	348	5.93	95.40
AMOTSCO	0	Número de pólizas de motocicleta = 0	-0.1539	332	5600	348	5.93	95.40
MSKB2	0	Clase social B2 = 0%	-0.1576	58	990	348	5.86	16.67
PLEVEN	6	Contribución de seguros de vida = f 1,000 – 4,999	-0.1857	2	38	348	5.26	0.57
MOPLMIDD	1	Educación de nivel medio = 1 - 10%	-0.1925	22	383	348	5.74	6.32
MAUT2	5	2 autos = 50 - 62%	-0.1958	3	56	348	5.36	0.86
ABRAND	2	Número de pólizas de incendio = 2	-0.1997	7	126	348	5.56	2.01
PGEZONG	0	Contribución de pólizas de seguro contra accidentes familiares = f 0	-0.2068	342	5784	348	5.91	98.28
AGEZONG	0	Número de pólizas de seguro contra accidentes familiares = 0	-0.2068	342	5784	348	5.91	98.28
MAUT2	3	2 autos = 24 - 36%	-0.2177	22	385	348	5.71	6.32
MINK123M	0	Ingresos > 123,000 = 0%	-0.2344	289	4900	348	5.90	83.05
MFWEKIND	4	Familia sin hijos = 37 - 49%	-0.2455	66	1137	348	5.80	18.97
MAANTHUI	5	Número de casas = 5	-0.2521	0	1	348	0.00	0.00
MAANTHUI	6	Número de casas = 6	-0.2521	0	1	348	0.00	0.00
MAANTHUI	8	Número de casas = 8	-0.2521	0	1	348	0.00	0.00
MAANTHUI	10	Número de casas = 10	-0.2521	0	1	348	0.00	0.00
MRELSA	7	Unión libre = 76 - 88%	-0.2521	0	1	348	0.00	0.00
MSKD	9	Clase social D = 100%	-0.2521	0	1	348	0.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
MAUT2	7	2 autos = 76 - 88%	-0.2521	0	1	348	0.00	0.00
MINK7512	7	Ingresos 75 - 122,000 = 76 - 88%	-0.2521	0	1	348	0.00	0.00
MINK123M	5	Ingresos > 123,000 = 50 - 62%	-0.2521	0	1	348	0.00	0.00
MINK123M	7	Ingresos > 123,000 = 76 - 88%	-0.2521	0	1	348	0.00	0.00
MINK123M	9	Ingresos > 123,000 = 100%	-0.2521	0	1	348	0.00	0.00
PWABEDR	5	Contribución de seguros contra daño a terceros (empresas) = f 500 – 999	-0.2521	0	1	348	0.00	0.00
PPERSAUT	4	Contribución de pólizas de auto = f 200 – 499	-0.2521	0	1	348	0.00	0.00
PVRAAUT	4	Contribución de pólizas de camión de carga = f 200 – 499	-0.2521	0	1	348	0.00	0.00
PVRAAUT	9	Contribución de pólizas de camión de carga = f 20,000 - ?	-0.2521	0	1	348	0.00	0.00
PAANHANG	4	Contribución de pólizas de tráiler = f 200 – 499	-0.2521	0	1	348	0.00	0.00
PAANHANG	5	Contribución de pólizas de tráiler = f 500 – 999	-0.2521	0	1	348	0.00	0.00
PBROM	6	Contribución de pólizas de ciclomotor = f 1,000 – 4,999	-0.2521	0	1	348	0.00	0.00
PLEVEN	8	Contribución de seguros de vida = f 10,000 - 19,999	-0.2521	0	1	348	0.00	0.00
PLEVEN	9	Contribución de seguros de vida = f 20,000 - ?	-0.2521	0	1	348	0.00	0.00
PPERSONG	5	Contribución de pólizas de seguro contra accidentes = f 500 – 999	-0.2521	0	1	348	0.00	0.00
PWAOREG	4	Contribución de pólizas de seguro contra discapacidades = f 200 – 499	-0.2521	0	1	348	0.00	0.00
PWAOREG	5	Contribución de pólizas de seguro contra discapacidades = f 500 – 999	-0.2521	0	1	348	0.00	0.00
PBRAND	8	Contribución de pólizas de incendio = f 10,000 - 19,999	-0.2521	0	1	348	0.00	0.00
PZEILPL	3	Contribución de pólizas de tablas de surfeo = f 100 – 199	-0.2521	0	1	348	0.00	0.00
PINBOED	5	Contribución de pólizas de seguro de propiedad = f 500 – 999	-0.2521	0	1	348	0.00	0.00
PINBOED	6	Contribución de pólizas de seguro de propiedad = f 1,000 – 4,999	-0.2521	0	1	348	0.00	0.00
PBYSTAND	5	Contribución de pólizas de seguridad social = 5	-0.2521	0	1	348	0.00	0.00
AWABEDR	5	Número de seguros contra daño a terceros (empresas) = 5	-0.2521	0	1	348	0.00	0.00
APERSAUT	6	Número de pólizas de auto = 6	-0.2521	0	1	348	0.00	0.00
APERSAUT	7	Número de pólizas de auto = 7	-0.2521	0	1	348	0.00	0.00
ABESAUT	4	Número de pólizas de camioneta repartidora = 4	-0.2521	0	1	348	0.00	0.00
AMOTSCO	8	Número de pólizas de motocicleta = 8	-0.2521	0	1	348	0.00	0.00
AVRAAUT	3	Número de pólizas de camión de carga = 3	-0.2521	0	1	348	0.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
AWERKT	6	Número de pólizas de máquinas para agricultura = 6	-0.2521	0	1	348	0.00	0.00
ALEVEN	8	Número de seguros de vida = 8	-0.2521	0	1	348	0.00	0.00
ABRAND	7	Número de pólizas de incendio = 7	-0.2521	0	1	348	0.00	0.00
AINBOED	2	Número de pólizas de seguro de propiedad = 2	-0.2521	0	1	348	0.00	0.00
ABYSTAND	2	Número de pólizas de seguridad social = 2	-0.2521	0	1	348	0.00	0.00
ABESAUT	1	Número de pólizas de camioneta repartidora = 1	-0.2607	2	40	348	5.00	0.57
MOSTYPE	10	Subtipo de cliente = Familias estables	-0.2833	9	165	348	5.45	2.59
MSKD	6	Clase social D = 63 - 75%	-0.2833	1	22	348	4.55	0.29
PLEVEN	0	Contribución de seguros de vida = f 0	-0.3112	325	5529	348	5.88	93.39
ALEVEN	0	Número de seguros de vida = 0	-0.3112	325	5529	348	5.88	93.39
PBRAND	5	Contribución de pólizas de incendio = f 500 – 999	-0.3132	8	149	348	5.37	2.30
MSKB1	5	Clase social B1 = 50 - 62%	-0.3163	4	78	348	5.13	1.15
MGODOV	4	Otra religión = 37 - 49%	-0.3268	7	132	348	5.30	2.01
PFIETS	0	Contribución de pólizas de bicicleta = f 0	-0.3479	333	5675	348	5.87	95.69
AFIETS	0	Número de pólizas de bicicleta = 0	-0.3479	333	5675	348	5.87	95.69
MINK3045	0	Ingresos 30 - 45,000 = 0%	-0.3510	26	465	348	5.59	7.47
MAANTHUI	4	Número de casas = 4	-0.3566	0	2	348	0.00	0.00
MSKB2	9	Clase social B2 = 100%	-0.3566	0	2	348	0.00	0.00
PMOTSCO	7	Contribución de pólizas de motocicleta = f 5,000 – 9,999	-0.3566	0	2	348	0.00	0.00
PPERSONG	6	Contribución de pólizas de seguro contra accidentes = f 1,000 – 4,999	-0.3566	0	2	348	0.00	0.00
PWAOREG	7	Contribución de pólizas de seguro contra discapacidades = f 5,000 – 9,999	-0.3566	0	2	348	0.00	0.00
PPLEZIER	5	Contribución de pólizas de bote = f 500 – 999	-0.3566	0	2	348	0.00	0.00
AVRAAUT	2	Número de pólizas de camión de carga = 2	-0.3566	0	2	348	0.00	0.00
AAANHANG	3	Número de pólizas de tráiler = 3	-0.3566	0	2	348	0.00	0.00
AWERKT	3	Número de pólizas de máquinas para agricultura = 3	-0.3566	0	2	348	0.00	0.00
ABRAND	5	Número de pólizas de incendio = 5	-0.3566	0	2	348	0.00	0.00
MOSTYPE	33	Subtipo de cliente = Familias grandes de clase baja	-0.3581	46	810	348	5.68	13.22
MZFONDS	7	Servicio nacional de salud = 76 - 88%	-0.3600	87	1511	348	5.76	25.00
MZPART	2	Segura privado de salud = 11 - 23%	-0.3600	87	1511	348	5.76	25.00
MRELOV	9	Otra relación = 100%	-0.3742	1	24	348	4.17	0.29
MFALLEEN	8	Soltero = 89 - 99%	-0.3742	1	24	348	4.17	0.29
MAUT2	0	2 autos = 0%	-0.3742	107	1854	348	5.77	30.75

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
MBERHOOG	2	Estatus alto = 11 - 23%	-0.4033	78	1364	348	5.72	22.41
MOPLMIDD	6	Educación de nivel medio = 63 - 75%	-0.4073	19	348	348	5.46	5.46
MAUT2	1	2 autos = 1 - 10%	-0.4125	84	1468	348	5.72	24.14
MOSTYPE	5	Subtipo de cliente = Seniors variados	-0.4338	2	45	348	4.44	0.57
MOSTYPE	19	Subtipo de cliente = Juventud suburbana	-0.4367	0	3	348	0.00	0.00
MGODRK	8	Católico romano = 89 - 99%	-0.4367	0	3	348	0.00	0.00
MBERBOER	7	Granjero = 76 - 88%	-0.4367	0	3	348	0.00	0.00
PWALAND	2	Contribución de seguros contra daño a terceros (agricultura) = f 50 – 99	-0.4367	0	3	348	0.00	0.00
PPERSAUT	8	Contribución de pólizas de auto = f 10,000 - 19,999	-0.4367	0	3	348	0.00	0.00
PBESAUT	7	Contribución de pólizas de camioneta repartidora = f 5,000 – 9,999	-0.4367	0	3	348	0.00	0.00
PWERKT	6	Contribución de pólizas de máquinas para agricultura = f 1,000 – 4,999	-0.4367	0	3	348	0.00	0.00
PLEVEN	7	Contribución de seguros de vida = f 5,000 – 9,999	-0.4367	0	3	348	0.00	0.00
PPERSONG	1	Contribución de pólizas de seguro contra accidentes = f 1 – 49	-0.4367	0	3	348	0.00	0.00
PPERSONG	4	Contribución de pólizas de seguro contra accidentes = f 200 – 499	-0.4367	0	3	348	0.00	0.00
PINBOED	4	Contribución de pólizas de seguro de propiedad = f 200 – 499	-0.4367	0	3	348	0.00	0.00
ABESAUT	3	Número de pólizas de camioneta repartidora = 3	-0.4367	0	3	348	0.00	0.00
ATRACTOR	3	Número de pólizas de tractor = 3	-0.4367	0	3	348	0.00	0.00
ABRAND	4	Número de pólizas de incendio = 4	-0.4367	0	3	348	0.00	0.00
MBERARBO	2	Trabajadores no capacitados = 11 - 23%	-0.4463	82	1439	348	5.70	23.56
MBERMIDD	2	Administración media = 11 - 23%	-0.4503	85	1491	348	5.70	24.43
MGODPR	8	Protestante = 89 - 99%	-0.4632	3	65	348	4.62	0.86
MOSHOOFD	8	Tipo principal de cliente = Familia con adultos	-0.4722	89	1563	348	5.69	25.57
MBERARBO	6	Trabajadores no capacitados = 63 - 75%	-0.4935	6	122	348	4.92	1.72
MBERHOOG	5	Estatus alto = 50 - 62%	-0.5035	13	249	348	5.22	3.74
MGODRK	9	Católico romano = 100%	-0.5043	0	4	348	0.00	0.00
MBERBOER	9	Granjero = 100%	-0.5043	0	4	348	0.00	0.00
MINK7512	8	Ingresos 75 - 122,000 = 89 - 99%	-0.5043	0	4	348	0.00	0.00
PWABEDR	6	Contribución de seguros contra daño a terceros (empresas) = f 1,000 – 4,999	-0.5043	0	4	348	0.00	0.00
PWERKT	2	Contribución de pólizas de	-0.5043	0	4	348	0.00	0.00

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
		máquinas para agricultura = f 50 – 99						
PPERSONG	3	Contribución de pólizas de seguro contra accidentes = f 100 – 199	-0.5043	0	4	348	0.00	0.00
ABESAUT	2	Número de pólizas de camioneta repartidora = 2	-0.5043	0	4	348	0.00	0.00
AAANHANG	2	Número de pólizas de tráiler = 2	-0.5043	0	4	348	0.00	0.00
AWAOREG	2	Número de pólizas de seguro contra discapacidades = 2	-0.5043	0	4	348	0.00	0.00
MSKC	3	Clase social C = 24 - 36%	-0.5306	61	1090	348	5.60	17.53
MSKC	9	Clase social C = 100%	-0.5348	6	124	348	4.84	1.72
MGODOV	5	Otra religión = 50 - 62%	-0.5370	1	28	348	3.57	0.29
MSKB1	2	Clase social B1 = 11 - 23%	-0.5570	101	1783	348	5.66	29.02
MSKA	2	Clase social A = 11 - 23%	-0.5616	67	1198	348	5.59	19.25
MOSTYPE	15	Subtipo de cliente = Seniors cosmopolitanos	-0.5638	0	5	348	0.00	0.00
MAANTHUI	7	Número de casas = 7	-0.5638	0	5	348	0.00	0.00
MGODGE	8	Sin religión = 89 - 99%	-0.5638	0	5	348	0.00	0.00
MBERBOER	8	Granjero = 89 - 99%	-0.5638	0	5	348	0.00	0.00
MSKB1	7	Clase social B1 = 76 - 88%	-0.5638	0	5	348	0.00	0.00
APERSAUT	4	Número de pólizas de auto = 4	-0.5638	0	5	348	0.00	0.00
MFGEKIND	9	Familia con hijos = 100%	-0.5745	1	29	348	3.45	0.29
ATRACTOR	2	Número de pólizas de tractor = 2	-0.5745	1	29	348	3.45	0.29
MOPLLAAG	8	Educación de nivel bajo = 89 - 99%	-0.5776	13	254	348	5.12	3.74
MGODPR	6	Protestante = 63 - 75%	-0.5806	39	714	348	5.46	11.21
MOSTYPE	34	Subtipo de cliente = Familias grandes con niños que trabajan	-0.5874	9	182	348	4.95	2.59
MGEMLEEF	6	Edad promedio = 70 - 80 años	-0.6109	1	30	348	3.33	0.29
PPLEZIER	0	Contribución de pólizas de bote = f 0	-0.6114	335	5789	348	5.79	96.26
APLEZIER	0	Número de pólizas de bote = 0	-0.6114	335	5789	348	5.79	96.26
MGODRK	7	Católico romano = 76 - 88%	-0.6176	0	6	348	0.00	0.00
MSKB2	7	Clase social B2 = 76 - 88%	-0.6176	0	6	348	0.00	0.00
PAANHANG	3	Contribución de pólizas de tráiler = f 100 – 199	-0.6176	0	6	348	0.00	0.00
PWERKT	3	Contribución de pólizas de máquinas para agricultura = f 100 – 199	-0.6176	0	6	348	0.00	0.00
PINBOED	3	Contribución de pólizas de seguro de propiedad = f 100 – 199	-0.6176	0	6	348	0.00	0.00
AWAPART	2	Número de seguros contra daño a terceros = 2	-0.6176	0	6	348	0.00	0.00
AVRAAUT	1	Número de pólizas de camión de carga = 1	-0.6176	0	6	348	0.00	0.00
ATRACTOR	4	Número de pólizas de tractor = 4	-0.6176	0	6	348	0.00	0.00
AWERKT	2	Número de pólizas de máquinas para agricultura = 2	-0.6176	0	6	348	0.00	0.00
PBYSTAND	0	Contribución de pólizas de	-0.6179	332	5740	348	5.78	95.40

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
		seguridad social = 0						
ABYSTAND	0	Número de pólizas de seguridad social = 0	-0.6179	332	5740	348	5.78	95.40
MINK4575	9	Ingresos 45 - 75,000 = 100%	-0.6365	4	91	348	4.40	1.15
MSKB2	1	Clase social B2 = 1 - 10%	-0.6418	47	861	348	5.46	13.51
APERSONG	1	Número de pólizas de seguro contra accidentes = 1	-0.6462	1	31	348	3.23	0.29
MOSTYPE	4	Subtipo de cliente = Departamentos con seniors afluentes	-0.6483	2	52	348	3.85	0.57
MFALLEEN	5	Soltero = 50 - 62%	-0.6504	13	259	348	5.02	3.74
MINK3045	7	Ingresos 30 - 45,000 = 76 - 88%	-0.6639	10	205	348	4.88	2.87
MSKB2	8	Clase social B2 = 89 - 99%	-0.6671	0	7	348	0.00	0.00
PWABEDR	1	Contribución de seguros contra daño a terceros (empresas) = f 1 – 49	-0.6671	0	7	348	0.00	0.00
PVRAAUT	6	Contribución de pólizas de camión de carga = f 1,000 – 4,999	-0.6671	0	7	348	0.00	0.00
ABRAND	3	Número de pólizas de incendio = 3	-0.6671	0	7	348	0.00	0.00
MRELSA	2	Unión libre = 11 - 23%	-0.6762	59	1075	348	5.49	16.95
MOPLHOOG	1	Educación de nivel alto = 1 - 10%	-0.6984	73	1322	348	5.52	20.98
MGODGE	9	Sin religión = 100%	-0.7132	0	8	348	0.00	0.00
MINK7512	6	Ingresos 75 - 122,000 = 63 - 75%	-0.7132	0	8	348	0.00	0.00
PWERKT	4	Contribución de pólizas de máquinas para agricultura = f 200 – 499	-0.7132	0	8	348	0.00	0.00
MGODRK	3	Católico romano = 24 - 36%	-0.7136	7	152	348	4.61	2.01
PBROM	2	Contribución de pólizas de ciclomotor = f 50 – 99	-0.7468	1	34	348	2.94	0.29
MGODPR	9	Protestante = 100%	-0.7507	8	173	348	4.62	2.30
ALEVEN	1	Número de seguros de vida = 1	-0.7507	8	173	348	4.62	2.30
MOSTYPE	17	Subtipo de cliente = Con maestría reciente en la ciudad	-0.7564	0	9	348	0.00	0.00
PLEVEN	1	Contribución de seguros de vida = f 1 – 49	-0.7564	0	9	348	0.00	0.00
PBRAND	7	Contribución de pólizas de incendio = f 5,000 – 9,999	-0.7564	0	9	348	0.00	0.00
MHHUUR	2	Casa rentada = 11 - 23%	-0.7652	38	717	348	5.30	10.92
MINKM30	7	Ingresos < 30,000 = 76 - 88%	-0.7851	7	156	348	4.49	2.01
PWALAND	3	Contribución de seguros contra daño a terceros (agricultura) = f 100 – 199	-0.7862	2	57	348	3.51	0.57
MOSTYPE	22	Subtipo de cliente = Habitantes variados en departamentos	-0.7916	4	98	348	4.08	1.15
MBERMIDD	0	Administración media = 0%	-0.7952	35	667	348	5.25	10.06
PBESAUT	5	Contribución de pólizas de camioneta repartidora = f 500 – 999	-0.7973	0	10	348	0.00	0.00
MINK123M	3	Ingresos > 123,000 = 24 - 36%	-0.8098	1	36	348	2.78	0.29
MSKB1	1	Clase social B1 = 1 - 10%	-0.8185	81	1480	348	5.47	23.28

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(C)	P(C|***X***)	P(***X***|C)
MFGEKIND	2	Familia con hijos = 11 - 23%	-0.8240	57	1060	348	5.38	16.38
MHKOOP	7	Casa propia = 76 - 88%	-0.8271	38	724	348	5.25	10.92
MZPART	3	Segura privado de salud = 24 - 36%	-0.8321	45	849	348	5.30	12.93
MRELSA	3	Unión libre = 24 - 36%	-0.8376	7	159	348	4.40	2.01
MKOOPKLA	5	Clase con poder adquisitivo = 50 - 62%	-0.8469	30	583	348	5.15	8.62
MSKC	7	Clase social C = 76 - 88%	-0.8507	10	217	348	4.61	2.87
MAUT1	6	1 auto = 63 - 75%	-0.8692	91	1663	348	5.47	26.15
AWERKT	1	Número de pólizas de máquinas para agricultura = 1	-0.8734	0	12	348	0.00	0.00
MGODGE	5	Sin religión = 50 - 62%	-0.8919	51	963	348	5.30	14.66
MBERBOER	3	Granjero = 24 - 36%	-0.8986	6	143	348	4.20	1.72
MSKC	4	Clase social C = 37 - 49%	-0.9017	62	1159	348	5.35	17.82
MRELSA	6	Unión libre = 63 - 75%	-0.9091	0	13	348	0.00	0.00
MAUT0	8	Sin auto = 89 - 99%	-0.9091	0	13	348	0.00	0.00
MHHUUR	4	Casa rentada = 37 - 49%	-0.9096	26	517	348	5.03	7.47
MHKOOP	4	Casa propia = 37 - 49%	-0.9115	25	499	348	5.01	7.18
ATRACTOR	1	Número de pólizas de tractor = 1	-0.9370	4	105	348	3.81	1.15
MHKOOP	5	Casa propia = 50 - 62%	-0.9401	26	520	348	5.00	7.47
MBERBOER	6	Granjero = 63 - 75%	-0.9434	0	14	348	0.00	0.00
MBERMIDD	8	Administración media = 89 - 99%	-0.9434	0	14	348	0.00	0.00
MAUT1	1	1 auto = 1 - 10%	-0.9434	0	14	348	0.00	0.00
ABROM	2	Número de pólizas de ciclomotor = 2	-0.9434	0	14	348	0.00	0.00
MGODPR	4	Protestante = 37 - 49%	-0.9529	87	1607	348	5.41	25.00
MRELGE	0	Casado = 0%	-0.9625	2	64	348	3.13	0.57
MSKB2	5	Clase social B2 = 50 - 62%	-0.9687	17	357	348	4.76	4.89
MOSTYPE	21	Subtipo de cliente = Juventud urbana sin dinero ni posesiones	-0.9765	0	15	348	0.00	0.00
MZFONDS	1	Servicio nacional de salud = 1 - 10%	-0.9765	0	15	348	0.00	0.00
MZPART	8	Segura privado de salud = 89 - 99%	-0.9765	0	15	348	0.00	0.00
MOSTYPE	16	Subtipo de cliente = Estudiantes en departamentos	-1.0085	0	16	348	0.00	0.00
MBERARBO	7	Trabajadores no capacitados = 76 - 88%	-1.0099	2	66	348	3.03	0.57
MSKB1	0	Clase social B1 = 0%	-1.0176	72	1353	348	5.32	20.69
MFWEKIND	5	Familia sin hijos = 50 - 62%	-1.0286	58	1106	348	5.24	16.67
MSKB1	9	Clase social B1 = 100%	-1.0396	0	17	348	0.00	0.00
MAUT0	9	Sin auto = 100%	-1.0396	0	17	348	0.00	0.00
PWABEDR	4	Contribución de seguros contra daño a terceros (empresas) = f 200 – 499	-1.0396	0	17	348	0.00	0.00
MZFONDS	6	Servicio nacional de salud = 63 - 75%	-1.0412	45	875	348	5.14	12.93
MBERZELF	0	Empresario = 0%	-1.0656	233	4171	348	5.59	66.95
MFGEKIND	5	Familia con hijos = 50 - 62%	-1.0663	30	606	348	4.95	8.62
MBERMIDD	1	Administración media = 1 - 10%	-1.0692	19	403	348	4.71	5.46

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
MRELSA	5	Unión libre = 50 - 62%	-1.0697	0	18	348	0.00	0.00
MGODPR	3	Protestante = 24 - 36%	-1.0882	29	590	348	4.92	8.33
MGODPR	1	Protestante = 1 - 10%	-1.0967	5	134	348	3.73	1.44
MOSTYPE	18	Subtipo de cliente = Juventud soltera	-1.0990	0	19	348	0.00	0.00
MAUT1	0	1 auto = 0%	-1.0990	0	19	348	0.00	0.00
MOPLMIDD	2	Educación de nivel medio = 11 - 23%	-1.1035	48	937	348	5.12	13.79
MBERARBO	5	Trabajadores no capacitados = 50 - 62%	-1.1094	15	331	348	4.53	4.31
MKOOPKLA	4	Clase con poder adquisitivo = 37 - 49%	-1.1117	46	902	348	5.10	13.22
MHHUUR	5	Casa rentada = 50 - 62%	-1.1151	25	519	348	4.82	7.18
MGODGE	4	Sin religión = 37 - 49%	-1.1246	70	1334	348	5.25	20.11
MGEMOMV	2	Habitantes promedio por casa = 2	-1.1310	115	2131	348	5.40	33.05
MOSTYPE	26	Subtipo de cliente = Adultos mayores con casa propia	-1.1380	1	48	348	2.08	0.29
MINKM30	8	Ingresos < 30,000 = 89 - 99%	-1.1380	1	48	348	2.08	0.29
MRELOV	8	Otra relación = 89 - 99%	-1.1554	0	21	348	0.00	0.00
MINKGEM	1	Ingreso promedio = 1 - 10%	-1.1624	1	49	348	2.04	0.29
PMOTSCO	6	Contribución de pólizas de motocicleta = f 1,000 – 4,999	-1.1624	1	49	348	2.04	0.29
MOSTYPE	9	Subtipo de cliente = Familias completas modernas	-1.1681	12	278	348	4.32	3.45
MRELGE	5	Casado = 50 - 62%	-1.1720	48	946	348	5.07	13.79
MINKGEM	9	Ingreso promedio = 100%	-1.1826	0	22	348	0.00	0.00
MGODRK	0	Católico romano = 0%	-1.1841	177	3228	348	5.48	50.86
MOSTYPE	30	Subtipo de cliente = Solteros mayores religiosos	-1.1856	4	118	348	3.39	1.15
MOSTYPE	27	Subtipo de cliente = Seniors en departamentos	-1.1863	1	50	348	2.00	0.29
PWAPART	1	Contribución de seguros contra daño a terceros = f 1 – 49	-1.1944	8	201	348	3.98	2.30
MFALLEEN	9	Soltero = 100%	-1.2092	0	23	348	0.00	0.00
MINK123M	4	Ingresos > 123,000 = 37 - 49%	-1.2352	0	24	348	0.00	0.00
MBERARBO	4	Trabajadores no capacitados = 37 - 49%	-1.2365	38	772	348	4.92	10.92
MOSTYPE	28	Subtipo de cliente = Adultos mayores residenciales	-1.2607	0	25	348	0.00	0.00
MAUT0	7	Sin auto = 76 - 88%	-1.2607	0	25	348	0.00	0.00
MINKGEM	0	Ingreso promedio = 0%	-1.2607	0	25	348	0.00	0.00
MBERARBO	9	Trabajadores no capacitados = 100%	-1.2857	0	26	348	0.00	0.00
PTRACTOR	3	Contribución de pólizas de tractor = f 100 – 199	-1.2919	2	79	348	2.53	0.57
PTRACTOR	4	Contribución de pólizas de tractor = f 200 – 499	-1.3101	0	27	348	0.00	0.00
MHHUUR	7	Casa rentada = 76 - 88%	-1.3103	19	425	348	4.47	5.46
MHKOOP	2	Casa propia = 11 - 23%	-1.3210	19	426	348	4.46	5.46
MBERARBG	5	Trabajadores capacitados = 50 - 62%	-1.3248	13	310	348	4.19	3.74
MBERARBG	6	Trabajadores capacitados = 63 - 75%	-1.3309	6	169	348	3.55	1.72

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
PLEVEN	2	Contribución de seguros de vida = f 50 – 99	-1.3342	0	28	348	0.00	0.00
MFALLEEN	6	Soltero = 63 - 75%	-1.3442	4	127	348	3.15	1.15
MOSTYPE	25	Subtipo de cliente = Seniors jóvenes en la ciudad	-1.3516	2	82	348	2.44	0.57
MAUT1	2	1 auto = 11 - 23%	-1.3663	1	58	348	1.72	0.29
MOSTYPE	35	Subtipo de cliente = Familias de pueblo	-1.3816	8	214	348	3.74	2.30
MBERBOER	5	Granjero = 50 - 62%	-1.3875	1	59	348	1.69	0.29
MFALLEEN	3	Soltero = 24 - 36%	-1.4033	41	848	348	4.83	11.78
PWALAND	4	Contribución de seguros contra daño a terceros (agricultura) = f 200 – 499	-1.4085	1	60	348	1.67	0.29
MFWEKIND	0	Familia sin hijos = 0%	-1.4136	5	153	348	3.27	1.44
MOSTYPE	29	Subtipo de cliente = Seniors sin patio delantero	-1.4285	2	86	348	2.33	0.57
MSKA	1	Clase social A = 1 - 10%	-1.4679	80	1569	348	5.10	22.99
MINK3045	5	Ingresos 30 - 45,000 = 50 - 62%	-1.4722	45	931	348	4.83	12.93
MRELOV	7	Otra relación = 76 - 88%	-1.4898	1	64	348	1.56	0.29
MOPLLAAG	5	Educación de nivel bajo = 50 - 62%	-1.5021	49	1009	348	4.86	14.08
MGODRK	4	Católico romano = 37 - 49%	-1.5291	1	66	348	1.52	0.29
MFALLEEN	7	Soltero = 76 - 88%	-1.5485	1	67	348	1.49	0.29
MAANTHUI	3	Número de casas = 3	-1.5746	0	39	348	0.00	0.00
MSKD	2	Clase social D = 11 - 23%	-1.5791	40	852	348	4.69	11.49
MBERHOOG	1	Estatus alto = 1 - 10%	-1.6041	61	1245	348	4.90	17.53
AWALAND	1	Número de seguros contra daño a terceros (agricultura) = 1	-1.6068	3	120	348	2.50	0.86
PPERSAUT	7	Contribución de pólizas de auto = f 5,000 – 9,999	-1.6145	0	41	348	0.00	0.00
MAUT1	5	1 auto = 50 - 62%	-1.6159	59	1210	348	4.88	16.95
MGEMLEEF	1	Edad promedio = 20 - 30 años	-1.6786	1	74	348	1.35	0.29
MRELGE	1	Casado = 1 - 10%	-1.6965	1	75	348	1.33	0.29
MZFONDS	9	Servicio nacional de salud = 100%	-1.7236	39	852	348	4.58	11.21
MZPART	0	Segura privado de salud = 0%	-1.7236	39	852	348	4.58	11.21
MBERMIDD	3	Administración media = 24 - 36%	-1.7313	68	1394	348	4.88	19.54
MBERBOER	4	Granjero = 37 - 49%	-1.7318	1	77	348	1.30	0.29
MBERBOER	2	Granjero = 11 - 23%	-1.7413	20	487	348	4.11	5.75
MGODPR	0	Protestante = 0%	-1.7492	1	78	348	1.28	0.29
MRELSA	4	Unión libre = 37 - 49%	-1.7492	1	78	348	1.28	0.29
MINKM30	9	Ingresos < 30,000 = 100%	-1.8006	0	51	348	0.00	0.00
MOSTYPE	24	Subtipo de cliente = Jóvenes con bajo nivel educativo	-1.8107	5	180	348	2.78	1.44
MOSHOOFD	4	Tipo principal de cliente = Solteros con carrera	-1.8182	0	52	348	0.00	0.00
MFWEKIND	1	Familia sin hijos = 1 - 10%	-1.8400	10	292	348	3.42	2.87
MOSTYPE	31	Subtipo de cliente = Católicos de bajo ingresos	-1.8424	6	205	348	2.93	1.72
MFALLEEN	4	Soltero = 37 - 49%	-1.8557	21	519	348	4.05	6.03

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|*C*)
MINKM30	3	Ingresos < 30,000 = 24 - 36%	-1.8614	50	1079	348	4.63	14.37
MOPLLAAG	7	Educación de nivel bajo = 76 - 88%	-1.8766	27	640	348	4.22	7.76
MSKA	9	Clase social A = 100%	-1.8868	0	56	348	0.00	0.00
MAUT1	3	1 auto = 24 - 36%	-1.8894	7	231	348	3.03	2.01
MAUT0	3	Sin auto = 24 - 36%	-1.9016	49	1066	348	4.60	14.08
MBERARBG	4	Trabajadores capacitados = 37 - 49%	-1.9057	25	604	348	4.14	7.18
MBERHOOG	0	Estatus alto = 0%	-1.9552	73	1524	348	4.79	20.98
PBROM	4	Contribución de pólizas de ciclomotor = f 200 – 499	-2.0013	0	63	348	0.00	0.00
MSKA	0	Clase social A = 0%	-2.0121	84	1738	348	4.83	24.14
MAUT0	5	Sin auto = 50 - 62%	-2.0468	4	174	348	2.30	1.15
MBERARBO	3	Trabajadores no capacitados = 24 - 36%	-2.0632	50	1109	348	4.51	14.37
MINKM30	6	Ingresos < 30,000 = 63 - 75%	-2.0980	9	293	348	3.07	2.59
MSKD	5	Clase social D = 50 - 62%	-2.0996	1	100	348	1.00	0.29
MRELOV	4	Otra relación = 37 - 49%	-2.1100	26	648	348	4.01	7.47
MRELOV	6	Otra relación = 63 - 75%	-2.1122	4	179	348	2.23	1.15
PBRAND	6	Contribución de pólizas de incendio = f 1,000 – 4,999	-2.1226	3	155	348	1.94	0.86
MOSTYPE	40	Subtipo de cliente = Granjas de familias grandes	-2.1245	0	71	348	0.00	0.00
MKOOPKLA	2	Clase con poder adquisitivo = 11 - 23%	-2.1287	15	425	348	3.53	4.31
MOSTYPE	41	Subtipo de cliente = Rurales variados	-2.1370	5	205	348	2.44	1.44
MRELGE	2	Casado = 11 - 23%	-2.1493	3	157	348	1.91	0.86
MINK4575	0	Ingresos 45 - 75,000 = 0%	-2.1562	38	891	348	4.26	10.92
MKOOPKLA	3	Clase con poder adquisitivo = 24 - 36%	-2.1713	71	1524	348	4.66	20.40
MBERBOER	1	Granjero = 1 - 10%	-2.1719	36	854	348	4.22	10.34
MRELGE	4	Casado = 37 - 49%	-2.1950	10	324	348	3.09	2.87
MZFONDS	8	Servicio nacional de salud = 89 - 99%	-2.1988	28	699	348	4.01	8.05
MZPART	1	Segura privado de salud = 1 - 10%	-2.1988	28	699	348	4.01	8.05
PBRAND	1	Contribución de pólizas de incendio = f 1 – 49	-2.2019	3	161	348	1.86	0.86
MHHUUR	9	Casa rentada = 100%	-2.2076	31	760	348	4.08	8.91
MHKOOP	0	Casa propia = 0%	-2.2076	31	760	348	4.08	8.91
MGEMOMV	1	Habitantes promedio por casa = 1	-2.2466	8	284	348	2.82	2.30
MOPLMIDD	0	Educación de nivel medio = 0%	-2.3143	14	423	348	3.31	4.02
MOSHOOFD	7	Tipo principal de cliente = Retirados y religiosos	-2.3158	20	550	348	3.64	5.75
MRELGE	3	Casado = 24 - 36%	-2.3410	6	246	348	2.44	1.72
MSKD	4	Clase social D = 37 - 49%	-2.3528	5	223	348	2.24	1.44
MOSHOOFD	6	Tipo principal de cliente = Seniors viajeros	-2.4316	4	205	348	1.95	1.15
MSKD	3	Clase social D = 24 - 36%	-2.4827	14	441	348	3.17	4.02
MINK4575	1	Ingresos 45 - 75,000 = 1 - 10%	-2.5131	24	657	348	3.65	6.90
MINKM30	4	Ingresos < 30,000 = 37 - 49%	-2.5515	21	599	348	3.51	6.03

Variable	Valor	Descripción	ε	N(C^***X***)	N(***X***)	N(*C*)	P(*C*|***X***)	P(***X***|C)
MRELOV	5	Otra relación = 50 - 62%	-2.5604	6	266	348	2.26	1.72
MBERARBG	3	Trabajadores capacitados = 24 - 36%	-2.5629	49	1167	348	4.20	14.08
MGODGE	6	Sin religión = 63 - 75%	-2.5688	4	217	348	1.84	1.15
MGODOV	1	Otra religión = 1 - 10%	-2.5739	93	2014	348	4.62	26.72
MOPLLAAG	9	Educación de nivel bajo = 100%	-2.6538	8	323	348	2.48	2.30
MINK4575	2	Ingresos 45 - 75,000 = 11 - 23%	-2.6739	48	1165	348	4.12	13.79
PBROM	3	Contribución de pólizas de ciclomotor = f 100 – 199	-2.7270	6	282	348	2.13	1.72
MAUT1	4	1 auto = 37 - 49%	-2.7459	13	448	348	2.90	3.74
MOPLLAAG	6	Educación de nivel bajo = 63 - 75%	-2.7633	32	856	348	3.74	9.20
MHKOOP	1	Casa propia = 1 - 10%	-2.8730	16	530	348	3.02	4.60
MHHUUR	8	Casa rentada = 89 - 99%	-2.8894	16	532	348	3.01	4.60
MOSHOOFD	10	Tipo principal de cliente = Granjeros	-2.9193	5	276	348	1.81	1.44
MOSTYPE	23	Subtipo de cliente = Jóvenes crecientes	-2.9296	4	251	348	1.59	1.15
MSKC	5	Clase social C = 50 - 62%	-2.9394	46	1168	348	3.94	13.22
MKOOPKLA	1	Clase con poder adquisitivo = 1 - 10%	-2.9749	18	587	348	3.07	5.17
MINKM30	5	Ingresos < 30,000 = 50 - 62%	-3.0002	17	568	348	2.99	4.89
MINK7512	0	Ingresos 75 - 122,000 = 0%	-3.1114	152	3246	348	4.68	43.68
MINKGEM	2	Ingreso promedio = 11 - 23%	-3.1267	20	651	348	3.07	5.75
ABROM	1	Número de pólizas de ciclomotor = 1	-3.2014	8	382	348	2.09	2.30
MOSHOOFD	5	Tipo principal de cliente = Buen nivel de vida	-3.3619	15	569	348	2.64	4.31
MOPLHOOG	0	Educación de nivel alto = 0%	-3.3987	91	2147	348	4.24	26.15
MAUT0	4	Sin auto = 37 - 49%	-3.8454	13	587	348	2.21	3.74
PPERSAUT	5	Contribución de pólizas de auto = f 500 – 999	-3.8574	14	613	348	2.28	4.02
PBRAND	0	Contribución de pólizas de incendio = f 0	-4.1138	109	2666	348	4.09	31.32
ABRAND	0	Número de pólizas de incendio = 0	-4.1138	109	2666	348	4.09	31.32
PWAPART	0	Contribución de seguros contra daño a terceros = f 0	-4.3699	147	3482	348	4.22	42.24
AWAPART	0	Número de seguros contra daño a terceros = 0	-4.3699	147	3482	348	4.22	42.24
MINKGEM	3	Ingreso promedio = 24 - 36%	-4.4608	69	1932	348	3.57	19.83
PBRAND	2	Contribución de pólizas de incendio = f 50 – 99	-4.7377	6	535	348	1.12	1.72
PPERSAUT	0	Contribución de pólizas de auto = f 0	-7.7546	72	2845	348	2.53	20.69
APERSAUT	0	Número de pólizas de auto = 0	-7.7546	72	2845	348	2.53	20.69

Tabla A.6 Valores de ε para los datos de ColL.

Apéndice B: Información sobre datos DxCG

Valor	Descripción
1	Male
2	Female

Tabla B.1 Descripciones de valores tipo T1.

Valor	Descripción
1	0 to 3 years
2	4 to 18 years
3	19 to 30 years
4	31 to 40 years
5	41 to 50 years
6	51 to 60 years
7	61 to 70 years
8	Greater than 70 years

Tabla B.2 Descripciones de valores tipo T2.

Valor	Descripción
1	0 to 3 months
2	4 to 6 months
3	7 to 9 months
4	10 to 12 months

Tabla B.3 Descripciones de valores tipo T3.

Valor	Descripción
1	Top 0.5%
2	Top 1.5%
3	Top 3%
4	Top 5%
5	Top 10%
6	Top 80%

Tabla B.4 Descripciones de valores tipo T4.

Valor	Descripción
1	$\leq$ -0.07
2	> –0.07 and $\leq$ 0.2
3	> 0.2

Tabla B.5 Descripciones de valores tipo T5.

Valor	Descripción
1	$\leq$ 1.0
2	>1.0

Tabla B.6 Descripciones de valores tipo T6.

Valor	Descripción
1	≤ -1
2	> -1 and ≤ 1
3	> 1

Tabla B.7 Descripciones de valores tipo T7.

Valor	Descripción
1	≤ 0.3
2	> 0.3 and ≤ 0.5
3	> 0.5 and ≤ 0.8
4	> 0.8 and ≤ 1.0
5	> 1.0

Tabla B.8 Descripciones de valores tipo T8.

Valor	Descripción
0	0 days
1	1 to 5 days
2	6 to 10 days
3	11 to 20 days
4	21 to 30 days
5	31 to 40 days
6	41 to 50 days
7	≥ 51 days

Tabla B.9 Descripciones de valores tipo T9.

Valor	Descripción
0	0 hospitalization
1	1 hospitalization
2	2 to 3 hospitalization
3	≥ 4 hospitalization

Tabla B.10 Descripciones de valores tipo T10.

Valor	Descripción
1	0 to 2 days
2	3 to 5 days
3	6 to 8 days
4	9 to 10 days
5	11 to 12 days
6	≥ 13 days

Tabla B.11 Descripciones de valores tipo T11.

Valor	Descripción
0	Decrease or not change of level
1	Increase of 1 level
2	Increase of 2 levels
3	Increase of 3 levels
4	Increase of 4 levels
5	Increase of 5 levels

Tabla B.12 Descripciones de valores tipo T12.

Valor	Descripción
1	0 to 4
2	5 to 10
3	11 to 20
4	21 to 30
5	31 to 40
6	$\geq$ 41

Tabla B.13 Descripciones de valores tipo T13.

Valor	Descripción
0	None
1	Insuline
2	Oral
3	Both

Tabla B.14 Descripciones de valores tipo T14.

Variable	Datos 1997 (Todos los datos)				Datos 1997 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
SEX	1	2	1.48	0.50	1	2	1.40	0.49
AGE	0	63	50.23	11.37	22	63	51.66	9.32
ELIG1	1	12	11.66	1.38	3	12	11.69	1.29
ELIG2	1	12	11.29	2.00	3	12	11.17	1.96
ACOVY	0.00	708608.28	7935.47	20675.81	436.39	589141.33	88958.68	108839.61
COVYI	0.00	687386.95	2896.30	15525.92	0.00	481217.52	42533.36	79164.01
COVYO	0.00	357990.10	3426.28	8366.76	338.50	357990.10	42166.10	58632.38
COVYD	0.00	44486.74	1612.89	2015.55	0.00	37170.27	4259.22	5134.30
X_1	0.00	109.02	0.82	2.06	0.00	22.98	1.89	2.96
X_2	0.00	76.69	0.87	2.05	0.00	15.06	1.93	2.49
X_3	0.00	58.38	0.61	1.69	0.00	13.20	1.40	2.09
X_4	0.00	99.26	1.35	2.73	0.02	42.27	3.86	5.83
SLOPE	-32.71	29.78	0.13	1.05	-6.85	11.66	0.54	1.88
OFFSET	-49.63	109.02	0.58	2.70	-17.13	23.37	0.93	4.25
SUDDENNESS	-1.60	1.20	0.25	0.50	-1.17	1.20	0.32	0.49
CONCAVITY	-2.00	2.00	-0.40	0.96	-2.00	1.93	-0.48	0.94
RMSE	-1.00	31.83	0.62	1.26	0.06	11.67	1.23	1.56
NLR	-1.00	1.67	0.60	0.34	0.05	1.48	0.59	0.30
COVYI_1	0.00	292010.66	709.77	6203.50	0.00	157251.90	12351.10	30441.19

Variable	Datos 1997 (Todos los datos)				Datos 1997 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
COVYI_2	0.00	439851.94	731.19	7140.85	0.00	152274.87	9244.16	23184.01
COVYI_3	0.00	219533.93	427.86	4242.06	0.00	117637.41	5197.31	14931.95
COVYI_4	0.00	657344.40	1035.49	8954.68	0.00	278934.00	15868.55	38885.23
TLOS	0	342	1.23	7.42	0	189	16.44	30.62
NHOSP	0	20	0.22	0.73	0	20	2.10	3.12
HOS1	0	164	0.31	3.07	0	68	4.71	12.36
HOS2	0	241	0.29	2.86	0	47	3.41	7.83
HOS3	0	294	0.19	2.92	0	63	1.99	6.52
HOS4	0	156	0.36	3.04	0	126	5.79	14.82
COVYO_1	0.00	93853.38	782.48	2484.63	0.00	78550.25	7496.84	12852.39
COVYO_2	0.00	159361.84	839.26	2789.94	0.00	110837.21	9841.10	16094.00
COVYO_3	0.00	117935.07	572.81	2036.29	0.00	85268.70	7219.91	12863.91
COVYO_4	0.00	242192.40	1240.89	3828.13	0.00	242192.40	17679.76	29927.32
COVYD_1	0.00	19620.26	348.84	525.04	0.00	12058.75	900.55	1510.55
COVYD_2	0.00	18155.07	378.21	537.22	0.00	7971.19	971.58	1208.83
COVYD_3	0.00	11627.42	273.54	419.79	0.00	9003.62	751.03	1054.29
COVYD_4	0.00	37723.78	616.06	867.52	0.00	11847.67	1645.63	2004.73
COSTRI4	-5	5	0.00	0.99	-5	4	0.12	1.83
COSTRI43	-5	5	0.00	0.99	-5	5	0.46	1.74
COSTRII4	-3	4	0.90	0.63	-3	4	0.45	1.49
COSTRII43	-3	3	0.00	0.46	-3	3	0.07	1.13
ACOVY_1	0.00	300273.48	1841.08	7352.89	0.00	188710.36	20748.49	36315.09
ACOVY_2	0.00	442044.46	1948.65	8335.70	0.00	171374.27	20056.85	30320.83
ACOVY_3	0.00	229408.00	1274.20	5096.33	0.00	127958.71	13168.25	20820.80
ACOVY_4	0.00	666657.05	2892.45	10700.13	11.16	308151.12	35193.94	51660.92
COSTDXCG	172.54	114690.58	7606.20	8677.07	2081.35	114690.58	37733.69	26656.37
SCORE	0.01	51.76	0.69	1.38	0.11	36.97	6.83	7.49
HCC001	0	1	0.00	0.01	0	0	0.00	0.00
HCC002	0	1	0.01	0.08	0	1	0.11	0.31
HCC003	0	1	0.00	0.06	0	1	0.01	0.08
HCC004	0	1	0.00	0.03	0	1	0.01	0.08
HCC005	0	1	0.00	0.04	0	1	0.02	0.14
HCC006	0	1	0.12	0.32	0	1	0.25	0.43
HCC007	0	1	0.01	0.08	0	1	0.06	0.24
HCC008	0	1	0.01	0.08	0	1	0.03	0.18
HCC009	0	1	0.01	0.10	0	1	0.06	0.23
HCC010	0	1	0.04	0.19	0	1	0.12	0.33
HCC011	0	1	0.00	0.06	0	1	0.01	0.08
HCC012	0	1	0.03	0.17	0	1	0.07	0.25
HCC013	0	1	0.06	0.23	0	1	0.10	0.30
HCC014	0	1	0.13	0.34	0	1	0.08	0.27
HCC015	0	1	0.02	0.13	0	1	0.30	0.46
HCC016	0	1	0.05	0.22	0	1	0.30	0.46
HCC017	0	1	0.05	0.21	0	1	0.21	0.41

Variable	Datos 1997 (Todos los datos)				Datos 1997 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
HCC018	0	1	0.08	0.27	0	1	0.27	0.45
HCC019	0	1	0.84	0.36	0	1	0.91	0.29
HCC020	0	1	0.21	0.41	0	1	0.50	0.50
HCC021	0	1	0.00	0.05	0	1	0.03	0.18
HCC022	0	1	0.01	0.12	0	1	0.08	0.27
HCC023	0	1	0.03	0.18	0	1	0.25	0.43
HCC024	0	1	0.32	0.47	0	1	0.37	0.48
HCC025	0	1	0.00	0.04	0	1	0.02	0.14
HCC026	0	1	0.00	0.06	0	1	0.03	0.18
HCC027	0	1	0.00	0.04	0	1	0.02	0.14
HCC028	0	1	0.00	0.04	0	1	0.01	0.08
HCC029	0	1	0.02	0.13	0	1	0.08	0.28
HCC030	0	1	0.01	0.11	0	1	0.06	0.23
HCC031	0	1	0.01	0.10	0	1	0.08	0.27
HCC032	0	1	0.01	0.09	0	1	0.05	0.22
HCC033	0	1	0.01	0.07	0	1	0.02	0.14
HCC034	0	1	0.04	0.19	0	1	0.12	0.33
HCC035	0	1	0.00	0.04	0	1	0.02	0.14
HCC036	0	1	0.17	0.38	0	1	0.50	0.50
HCC037	0	1	0.01	0.10	0	1	0.10	0.30
HCC038	0	1	0.02	0.14	0	1	0.08	0.27
HCC039	0	1	0.06	0.23	0	1	0.08	0.27
HCC040	0	1	0.02	0.15	0	1	0.01	0.08
HCC041	0	1	0.03	0.17	0	1	0.13	0.34
HCC042	0	1	0.00	0.02	0	0	0.00	0.00
HCC043	0	1	0.36	0.48	0	1	0.48	0.50
HCC044	0	1	0.00	0.06	0	1	0.06	0.24
HCC045	0	1	0.00	0.06	0	1	0.04	0.20
HCC046	0	1	0.01	0.12	0	1	0.11	0.31
HCC047	0	1	0.07	0.26	0	1	0.39	0.49
HCC048	0	1	0.00	0.06	0	1	0.03	0.16
HCC049	0	1	0.00	0.06	0	0	0.00	0.00
HCC050	0	1	0.00	0.07	0	1	0.03	0.18
HCC051	0	1	0.00	0.03	0	0	0.00	0.00
HCC052	0	1	0.00	0.06	0	1	0.01	0.08
HCC053	0	1	0.00	0.04	0	1	0.01	0.08
HCC054	0	1	0.00	0.05	0	0	0.00	0.00
HCC055	0	1	0.03	0.18	0	1	0.05	0.22
HCC056	0	1	0.00	0.04	0	1	0.01	0.08
HCC057	0	1	0.00	0.05	0	1	0.01	0.08
HCC058	0	1	0.03	0.17	0	1	0.03	0.18
HCC059	0	1	0.01	0.10	0	1	0.02	0.14
HCC060	0	1	0.04	0.20	0	1	0.08	0.28
HCC061	0	1	0.00	0.01	0	0	0.00	0.00

Variable	Datos 1997 (Todos los datos)				Datos 1997 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
HCC062	0	1	0.00	0.01	0	1	0.01	0.08
HCC063	0	1	0.00	0.01	0	0	0.00	0.00
HCC064	0	1	0.00	0.02	0	0	0.00	0.00
HCC065	0	1	0.00	0.03	0	0	0.00	0.00
HCC066	0	1	0.00	0.05	0	0	0.00	0.00
HCC067	0	1	0.00	0.03	0	1	0.01	0.08
HCC068	0	1	0.00	0.02	0	0	0.00	0.00
HCC069	0	1	0.00	0.05	0	1	0.01	0.08
HCC070	0	1	0.00	0.01	0	0	0.00	0.00
HCC071	0	1	0.02	0.15	0	1	0.13	0.34
HCC072	0	1	0.00	0.05	0	0	0.00	0.00
HCC073	0	1	0.00	0.03	0	0	0.00	0.00
HCC074	0	1	0.01	0.10	0	1	0.06	0.23
HCC075	0	1	0.00	0.03	0	1	0.01	0.12
HCC076	0	1	0.06	0.23	0	1	0.08	0.27
HCC077	0	1	0.00	0.02	0	0	0.00	0.00
HCC078	0	1	0.00	0.03	0	1	0.01	0.08
HCC079	0	1	0.01	0.11	0	1	0.09	0.29
HCC080	0	1	0.04	0.20	0	1	0.25	0.43
HCC081	0	1	0.01	0.09	0	1	0.03	0.18
HCC082	0	1	0.04	0.18	0	1	0.10	0.31
HCC083	0	1	0.04	0.19	0	1	0.12	0.32
HCC084	0	1	0.11	0.31	0	1	0.26	0.44
HCC085	0	1	0.00	0.05	0	1	0.03	0.16
HCC086	0	1	0.03	0.17	0	1	0.11	0.31
HCC087	0	1	0.00	0.02	0	0	0.00	0.00
HCC088	0	1	0.00	0.06	0	1	0.03	0.16
HCC089	0	1	0.00	0.06	0	1	0.07	0.25
HCC090	0	1	0.02	0.15	0	1	0.03	0.18
HCC091	0	1	0.31	0.46	0	1	0.44	0.50
HCC092	0	1	0.02	0.14	0	1	0.06	0.24
HCC093	0	1	0.04	0.19	0	1	0.12	0.32
HCC094	0	1	0.03	0.17	0	1	0.18	0.38
HCC095	0	1	0.00	0.05	0	1	0.01	0.08
HCC096	0	1	0.01	0.12	0	1	0.07	0.25
HCC097	0	1	0.02	0.15	0	1	0.08	0.28
HCC098	0	1	0.00	0.06	0	0	0.00	0.00
HCC099	0	1	0.00	0.04	0	1	0.01	0.12
HCC100	0	1	0.00	0.05	0	1	0.01	0.08
HCC101	0	1	0.00	0.03	0	0	0.00	0.00
HCC102	0	1	0.00	0.03	0	1	0.01	0.08
HCC103	0	1	0.00	0.05	0	1	0.01	0.12
HCC104	0	1	0.01	0.12	0	1	0.12	0.32
HCC105	0	1	0.04	0.19	0	1	0.21	0.41

Variable	Datos 1997 (Todos los datos)				Datos 1997 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
HCC106	0	1	0.05	0.21	0	1	0.19	0.40
HCC107	0	1	0.00	0.02	0	1	0.02	0.14
HCC108	0	1	0.03	0.18	0	1	0.07	0.25
HCC109	0	1	0.02	0.13	0	1	0.10	0.30
HCC110	0	1	0.04	0.19	0	1	0.04	0.20
HCC111	0	1	0.00	0.04	0	1	0.03	0.18
HCC112	0	1	0.00	0.06	0	1	0.03	0.16
HCC113	0	1	0.04	0.19	0	1	0.24	0.43
HCC114	0	1	0.01	0.10	0	1	0.08	0.28
HCC115	0	1	0.12	0.33	0	1	0.25	0.43
HCC116	0	1	0.00	0.02	0	0	0.00	0.00
HCC117	0	1	0.00	0.04	0	1	0.01	0.08
HCC118	0	1	0.00	0.06	0	1	0.03	0.16
HCC119	0	1	0.02	0.15	0	1	0.18	0.38
HCC120	0	1	0.05	0.22	0	1	0.20	0.40
HCC121	0	1	0.01	0.10	0	1	0.04	0.20
HCC122	0	1	0.03	0.17	0	1	0.05	0.22
HCC123	0	1	0.03	0.17	0	1	0.10	0.31
HCC124	0	1	0.07	0.26	0	1	0.11	0.31
HCC125	0	1	0.01	0.07	0	1	0.01	0.08
HCC126	0	1	0.01	0.11	0	1	0.03	0.16
HCC127	0	1	0.29	0.45	0	1	0.37	0.49
HCC128	0	1	0.00	0.06	0	1	0.07	0.25
HCC129	0	0	0.00	0.00	0	0	0.00	0.00
HCC130	0	1	0.00	0.05	0	1	0.12	0.32
HCC131	0	1	0.01	0.12	0	1	0.39	0.49
HCC132	0	1	0.01	0.07	0	1	0.07	0.25
HCC133	0	1	0.02	0.16	0	1	0.09	0.29
HCC134	0	1	0.01	0.08	0	1	0.02	0.14
HCC135	0	1	0.09	0.29	0	1	0.17	0.38
HCC136	0	1	0.04	0.20	0	1	0.29	0.46
HCC137	0	1	0.00	0.05	0	0	0.00	0.00
HCC138	0	1	0.02	0.15	0	1	0.02	0.14
HCC139	0	1	0.12	0.33	0	1	0.08	0.28
HCC140	0	1	0.08	0.27	0	1	0.07	0.25
HCC141	0	1	0.00	0.02	0	0	0.00	0.00
HCC142	0	1	0.00	0.04	0	0	0.00	0.00
HCC143	0	1	0.00	0.06	0	0	0.00	0.00
HCC144	0	1	0.01	0.12	0	0	0.00	0.00
HCC145	0	1	0.02	0.12	0	1	0.01	0.08
HCC146	0	1	0.02	0.13	0	0	0.00	0.00
HCC147	0	1	0.02	0.15	0	1	0.01	0.08
HCC148	0	1	0.00	0.06	0	1	0.03	0.18
HCC149	0	1	0.02	0.14	0	1	0.16	0.37

Variable	Datos 1997 (Todos los datos)				Datos 1997 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
HCC150	0	0	0.00	0.00	0	0	0.00	0.00
HCC151	0	1	0.00	0.01	0	0	0.00	0.00
HCC152	0	1	0.06	0.24	0	1	0.23	0.42
HCC153	0	1	0.18	0.38	0	1	0.26	0.44
HCC154	0	1	0.00	0.01	0	1	0.01	0.08
HCC155	0	1	0.00	0.07	0	0	0.00	0.00
HCC156	0	1	0.00	0.03	0	1	0.01	0.08
HCC157	0	1	0.00	0.05	0	0	0.00	0.00
HCC158	0	1	0.00	0.06	0	1	0.02	0.14
HCC159	0	1	0.01	0.10	0	1	0.01	0.12
HCC160	0	1	0.00	0.04	0	1	0.01	0.08
HCC161	0	1	0.00	0.05	0	1	0.06	0.23
HCC162	0	1	0.17	0.38	0	1	0.23	0.42
HCC163	0	1	0.03	0.17	0	1	0.10	0.31
HCC164	0	1	0.01	0.12	0	1	0.23	0.43
HCC165	0	1	0.01	0.11	0	1	0.09	0.29
HCC166	0	1	0.29	0.45	0	1	0.67	0.47
HCC167	0	1	0.34	0.47	0	1	0.67	0.47
HCC168	0	0	0.00	0.00	0	0	0.00	0.00
HCC169	0	0	0.00	0.00	0	0	0.00	0.00
HCC170	0	1	0.00	0.01	0	0	0.00	0.00
HCC171	0	1	0.00	0.01	0	0	0.00	0.00
HCC172	0	1	0.00	0.01	0	0	0.00	0.00
HCC173	0	0	0.00	0.00	0	0	0.00	0.00
HCC174	0	1	0.00	0.04	0	1	0.03	0.18
HCC175	0	1	0.00	0.04	0	1	0.03	0.16
HCC176	0	1	0.00	0.05	0	1	0.03	0.16
HCC177	0	1	0.00	0.04	0	1	0.02	0.14
HCC178	0	0	0.00	0.00	0	0	0.00	0.00
HCC179	0	1	0.04	0.19	0	1	0.23	0.43
HCC180	0	1	0.00	0.03	0	1	0.01	0.08
HCC181	0	1	0.00	0.05	0	1	0.03	0.16
HCC182	0	1	0.01	0.10	0	1	0.04	0.20
HCC183	0	1	0.33	0.47	0	1	0.44	0.50
HCC184	0	1	0.03	0.18	0	1	0.09	0.29
COMORB	0	48	6.83	4.92	1	48	16.96	9.33
RXTYPE	0	3	1.15	0.98	0	3	0.90	0.97
ACOVNY	0.00	1740070.19	7645.73	22610.25	123349.67	1740070.19	222240.03	168531.67

Tabla B.15 Principales estadísticas de los datos de 1997 (todos los datos y datos de la clase).

Variable	Datos 1998 (Todos los datos)				Datos 1998 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
SEX	1	2	1.46	0.50	1	2	1.43	0.50
AGE	0	95	51.61	11.63	19	87	54.39	9.56
ELIG1	1	12	11.48	1.72	3	12	11.69	1.27
ELIG2	1	12	11.36	2.00	3	12	11.28	1.81
ACOVY	0.00	5897911.92	8800.28	38914.07	283.29	582017.95	81473.60	100045.59
COVYI	0.00	5797069.24	3172.29	35143.16	0.00	375084.93	35166.48	62375.90
COVYO	0.00	439418.79	3698.57	9968.60	0.00	439418.79	42259.73	64773.57
COVYD	0.00	81601.97	1929.42	2503.12	0.00	24929.35	4047.39	3792.99
X_1	0.00	397.50	0.91	3.28	0.00	15.95	1.83	2.75
X_2	0.00	194.65	0.93	2.31	0.00	30.73	2.12	3.54
X_3	0.00	98.42	0.65	1.82	0.00	15.48	1.50	2.42
X_4	0.00	184.30	1.39	3.30	0.00	17.53	2.72	3.16
SLOPE	-119.09	55.29	0.12	1.41	-3.79	3.86	0.20	1.04
OFFSET	-92.15	397.35	0.67	3.85	-5.27	17.43	1.53	3.30
SUDDENNESS	-1.60	1.20	0.23	0.48	-1.41	1.20	0.22	0.50
CONCAVITY	-2.00	2.00	-0.37	0.93	-2.00	1.98	-0.35	0.93
RMSE	-1.00	108.85	0.63	1.58	0.02	12.49	1.10	1.46
NLR	-1.00	1.67	0.57	0.34	0.09	1.66	0.56	0.32
COVYI_1	0.00	5659754.24	894.36	30340.36	0.00	140058.24	7340.95	21311.33
COVYI_2	0.00	368339.40	723.26	6959.52	0.00	245515.14	9247.42	26121.67
COVYI_3	0.00	1535948.05	546.55	9721.17	0.00	353373.53	8031.95	33690.47
COVYI_4	0.00	314845.12	1004.88	7924.56	0.00	209490.70	10456.76	25045.18
TLOS	0	331	1.20	7.05	0	214	16.16	32.08
NHOSP	0	17	0.21	0.70	0	17	1.90	2.53
HOS1	0	179	0.30	2.97	0	76	3.73	11.82
HOS2	0	207	0.29	2.88	0	140	4.56	13.72
HOS3	0	154	0.18	2.23	0	154	2.85	12.87
HOS4	0	120	0.35	2.97	0	108	4.42	11.35
COVYO_1	0.00	131160.61	856.11	2914.05	0.00	131160.61	8198.01	17144.83
COVYO_2	0.00	139908.60	924.90	3209.09	0.00	124843.68	10274.93	17445.32
COVYO_3	0.00	158598.80	634.55	2533.74	0.00	95528.20	8172.39	14217.77
COVYO_4	0.00	155686.88	1273.39	4218.98	0.00	153199.75	15421.13	24871.86
COVYD_1	0.00	15554.78	426.26	626.66	0.00	9499.77	992.29	1249.58
COVYD_2	0.00	63496.32	466.35	746.94	0.00	7365.02	1012.59	1058.09
COVYD_3	0.00	28107.01	327.11	535.23	0.00	3521.06	601.45	603.55
COVYD_4	0.00	23325.86	702.61	967.85	0.00	9218.28	1430.81	1437.59
COSTRI4	-5	5	0.00	0.96	-5	5	-0.07	1.76
COSTRI43	-5	5	0.00	0.97	-5	5	0.28	1.75
COSTRII4	-3	3	0.00	0.46	-3	3	-0.09	1.22
COSTRII43	-3	3	0.00	0.46	-3	3	0.07	1.20
ACOVY_1	0.00	5670121.52	2176.73	30668.63	0.00	145299.24	16531.26	27727.98
ACOVY_2	0.00	382905.20	2114.51	8371.49	0.00	250566.81	20534.94	32617.79
ACOVY_3	0.00	1556780.76	1508.20	10501.50	0.00	359562.70	16805.79	38779.41

Variable	Datos 1998 (Todos los datos)				Datos 1998 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
ACOVY_4	0.00	353998.99	2980.87	10266.21	0.00	225273.67	27308.70	38057.98
COSTDXCG	172.54	143252.93	7949.74	8991.65	984.98	96740.28	37550.98	25827.90
SCORE	0.01	110.29	0.69	1.52	0.03	41.65	5.86	6.68
HCC001	0	1	0.00	0.02	0	1	0.01	0.07
HCC002	0	1	0.01	0.08	0	1	0.07	0.25
HCC003	0	1	0.00	0.05	0	1	0.01	0.10
HCC004	0	1	0.00	0.03	0	0	0.00	0.00
HCC005	0	1	0.00	0.03	0	1	0.02	0.12
HCC006	0	1	0.11	0.32	0	1	0.22	0.41
HCC007	0	1	0.01	0.09	0	1	0.04	0.20
HCC008	0	1	0.01	0.08	0	1	0.04	0.19
HCC009	0	1	0.01	0.10	0	1	0.07	0.25
HCC010	0	1	0.04	0.20	0	1	0.07	0.26
HCC011	0	1	0.00	0.06	0	1	0.02	0.14
HCC012	0	1	0.04	0.18	0	1	0.09	0.28
HCC013	0	1	0.06	0.23	0	1	0.07	0.25
HCC014	0	1	0.12	0.32	0	1	0.07	0.26
HCC015	0	1	0.02	0.14	0	1	0.27	0.44
HCC016	0	1	0.06	0.23	0	1	0.28	0.45
HCC017	0	1	0.05	0.21	0	1	0.24	0.43
HCC018	0	1	0.09	0.28	0	1	0.32	0.47
HCC019	0	1	0.85	0.36	0	1	0.89	0.31
HCC020	0	1	0.20	0.40	0	1	0.52	0.50
HCC021	0	1	0.00	0.04	0	1	0.02	0.12
HCC022	0	1	0.01	0.12	0	1	0.07	0.25
HCC023	0	1	0.03	0.17	0	1	0.19	0.39
HCC024	0	1	0.33	0.47	0	1	0.38	0.49
HCC025	0	1	0.00	0.04	0	1	0.01	0.10
HCC026	0	1	0.00	0.06	0	1	0.03	0.17
HCC027	0	1	0.00	0.05	0	1	0.01	0.10
HCC028	0	1	0.00	0.03	0	1	0.02	0.12
HCC029	0	1	0.02	0.14	0	1	0.06	0.23
HCC030	0	1	0.01	0.11	0	1	0.05	0.22
HCC031	0	1	0.01	0.10	0	1	0.08	0.28
HCC032	0	1	0.01	0.09	0	1	0.03	0.17
HCC033	0	1	0.01	0.07	0	1	0.02	0.14
HCC034	0	1	0.04	0.20	0	1	0.11	0.31
HCC035	0	1	0.00	0.04	0	1	0.01	0.07
HCC036	0	1	0.18	0.38	0	1	0.41	0.49
HCC037	0	1	0.01	0.10	0	1	0.07	0.26
HCC038	0	1	0.02	0.14	0	1	0.03	0.16
HCC039	0	1	0.06	0.23	0	1	0.13	0.34
HCC040	0	1	0.03	0.16	0	1	0.04	0.19
HCC041	0	1	0.04	0.19	0	1	0.10	0.30

Variable	Datos 1998 (Todos los datos)				Datos 1998 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
HCC042	0	1	0.00	0.03	0	0	0.00	0.00
HCC043	0	1	0.36	0.48	0	1	0.53	0.50
HCC044	0	1	0.00	0.05	0	1	0.04	0.19
HCC045	0	1	0.00	0.06	0	1	0.03	0.17
HCC046	0	1	0.02	0.13	0	1	0.07	0.25
HCC047	0	1	0.07	0.25	0	1	0.32	0.47
HCC048	0	1	0.00	0.06	0	1	0.04	0.20
HCC049	0	1	0.00	0.07	0	1	0.01	0.10
HCC050	0	1	0.01	0.07	0	1	0.04	0.19
HCC051	0	1	0.00	0.03	0	1	0.01	0.07
HCC052	0	1	0.00	0.07	0	1	0.01	0.07
HCC053	0	1	0.00	0.06	0	1	0.01	0.10
HCC054	0	1	0.00	0.05	0	0	0.00	0.00
HCC055	0	1	0.03	0.18	0	1	0.12	0.32
HCC056	0	1	0.00	0.05	0	1	0.03	0.16
HCC057	0	1	0.00	0.05	0	1	0.01	0.10
HCC058	0	1	0.03	0.17	0	1	0.05	0.22
HCC059	0	1	0.01	0.11	0	1	0.04	0.20
HCC060	0	1	0.05	0.21	0	1	0.07	0.25
HCC061	0	1	0.00	0.01	0	0	0.00	0.00
HCC062	0	1	0.00	0.01	0	0	0.00	0.00
HCC063	0	1	0.00	0.01	0	0	0.00	0.00
HCC064	0	1	0.00	0.02	0	0	0.00	0.00
HCC065	0	1	0.00	0.03	0	0	0.00	0.00
HCC066	0	1	0.00	0.05	0	0	0.00	0.00
HCC067	0	1	0.00	0.03	0	0	0.00	0.00
HCC068	0	1	0.00	0.02	0	0	0.00	0.00
HCC069	0	1	0.00	0.06	0	1	0.01	0.10
HCC070	0	1	0.00	0.02	0	1	0.01	0.07
HCC071	0	1	0.02	0.15	0	1	0.17	0.38
HCC072	0	1	0.00	0.05	0	1	0.02	0.12
HCC073	0	1	0.00	0.03	0	0	0.00	0.00
HCC074	0	1	0.01	0.08	0	1	0.03	0.16
HCC075	0	1	0.00	0.03	0	1	0.01	0.07
HCC076	0	1	0.05	0.23	0	1	0.12	0.32
HCC077	0	1	0.00	0.02	0	0	0.00	0.00
HCC078	0	1	0.00	0.03	0	0	0.00	0.00
HCC079	0	1	0.01	0.11	0	1	0.10	0.30
HCC080	0	1	0.05	0.21	0	1	0.29	0.46
HCC081	0	1	0.01	0.09	0	1	0.07	0.26
HCC082	0	1	0.03	0.18	0	1	0.12	0.33
HCC083	0	1	0.04	0.19	0	1	0.11	0.32
HCC084	0	1	0.11	0.32	0	1	0.36	0.48
HCC085	0	1	0.00	0.05	0	1	0.02	0.12

Variable	Datos 1998 (Todos los datos)				Datos 1998 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
HCC086	0	1	0.03	0.18	0	1	0.18	0.38
HCC087	0	1	0.00	0.02	0	0	0.00	0.00
HCC088	0	1	0.00	0.06	0	1	0.03	0.16
HCC089	0	1	0.00	0.06	0	1	0.06	0.23
HCC090	0	1	0.02	0.15	0	1	0.03	0.16
HCC091	0	1	0.33	0.47	0	1	0.50	0.50
HCC092	0	1	0.03	0.16	0	1	0.12	0.33
HCC093	0	1	0.04	0.19	0	1	0.12	0.33
HCC094	0	1	0.03	0.17	0	1	0.13	0.34
HCC095	0	1	0.00	0.06	0	1	0.02	0.14
HCC096	0	1	0.01	0.12	0	1	0.11	0.31
HCC097	0	1	0.02	0.15	0	1	0.10	0.30
HCC098	0	1	0.00	0.06	0	1	0.02	0.12
HCC099	0	1	0.00	0.05	0	1	0.01	0.07
HCC100	0	1	0.00	0.05	0	1	0.03	0.16
HCC101	0	1	0.00	0.03	0	0	0.00	0.00
HCC102	0	1	0.00	0.03	0	0	0.00	0.00
HCC103	0	1	0.00	0.05	0	1	0.02	0.12
HCC104	0	1	0.01	0.12	0	1	0.11	0.31
HCC105	0	1	0.04	0.19	0	1	0.26	0.44
HCC106	0	1	0.05	0.21	0	1	0.24	0.43
HCC107	0	1	0.00	0.02	0	1	0.01	0.07
HCC108	0	1	0.04	0.19	0	1	0.08	0.28
HCC109	0	1	0.02	0.13	0	1	0.07	0.26
HCC110	0	1	0.04	0.19	0	1	0.06	0.23
HCC111	0	1	0.00	0.04	0	1	0.03	0.16
HCC112	0	1	0.00	0.05	0	1	0.02	0.12
HCC113	0	1	0.04	0.19	0	1	0.20	0.40
HCC114	0	1	0.01	0.10	0	1	0.08	0.28
HCC115	0	1	0.12	0.32	0	1	0.22	0.42
HCC116	0	1	0.00	0.02	0	1	0.01	0.07
HCC117	0	1	0.00	0.04	0	0	0.00	0.00
HCC118	0	1	0.00	0.06	0	1	0.02	0.14
HCC119	0	1	0.02	0.15	0	1	0.20	0.40
HCC120	0	1	0.06	0.24	0	1	0.22	0.41
HCC121	0	1	0.01	0.11	0	1	0.04	0.20
HCC122	0	1	0.04	0.19	0	1	0.08	0.27
HCC123	0	1	0.04	0.19	0	1	0.13	0.34
HCC124	0	1	0.08	0.27	0	1	0.06	0.24
HCC125	0	1	0.01	0.07	0	0	0.00	0.00
HCC126	0	1	0.01	0.12	0	1	0.03	0.16
HCC127	0	1	0.26	0.44	0	1	0.26	0.44
HCC128	0	1	0.00	0.06	0	1	0.04	0.19
HCC129	0	0	0.00	0.00	0	0	0.00	0.00

Variable	Datos 1998 (Todos los datos)				Datos 1998 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
HCC130	0	1	0.00	0.05	0	1	0.13	0.34
HCC131	0	1	0.02	0.13	0	1	0.42	0.49
HCC132	0	1	0.01	0.08	0	1	0.10	0.30
HCC133	0	1	0.03	0.16	0	1	0.10	0.30
HCC134	0	1	0.01	0.08	0	1	0.01	0.10
HCC135	0	1	0.08	0.27	0	1	0.19	0.39
HCC136	0	1	0.04	0.20	0	1	0.34	0.48
HCC137	0	1	0.00	0.06	0	0	0.00	0.00
HCC138	0	1	0.02	0.15	0	1	0.02	0.14
HCC139	0	1	0.11	0.32	0	1	0.06	0.24
HCC140	0	1	0.08	0.27	0	1	0.11	0.32
HCC141	0	1	0.00	0.02	0	0	0.00	0.00
HCC142	0	1	0.00	0.03	0	0	0.00	0.00
HCC143	0	1	0.00	0.05	0	1	0.01	0.07
HCC144	0	1	0.01	0.10	0	1	0.01	0.10
HCC145	0	1	0.01	0.12	0	1	0.01	0.07
HCC146	0	1	0.01	0.12	0	1	0.01	0.07
HCC147	0	1	0.02	0.13	0	1	0.01	0.07
HCC148	0	1	0.00	0.07	0	1	0.06	0.24
HCC149	0	1	0.02	0.15	0	1	0.16	0.37
HCC150	0	1	0.00	0.01	0	0	0.00	0.00
HCC151	0	1	0.00	0.02	0	0	0.00	0.00
HCC152	0	1	0.07	0.25	0	1	0.21	0.41
HCC153	0	1	0.19	0.39	0	1	0.30	0.46
HCC154	0	1	0.00	0.01	0	0	0.00	0.00
HCC155	0	1	0.00	0.06	0	1	0.01	0.10
HCC156	0	1	0.00	0.03	0	0	0.00	0.00
HCC157	0	1	0.00	0.05	0	1	0.01	0.07
HCC158	0	1	0.00	0.05	0	1	0.01	0.10
HCC159	0	1	0.01	0.10	0	1	0.04	0.19
HCC160	0	1	0.00	0.04	0	0	0.00	0.00
HCC161	0	1	0.00	0.05	0	1	0.04	0.20
HCC162	0	1	0.17	0.37	0	1	0.30	0.46
HCC163	0	1	0.03	0.17	0	1	0.05	0.22
HCC164	0	1	0.02	0.12	0	1	0.16	0.37
HCC165	0	1	0.01	0.11	0	1	0.05	0.22
HCC166	0	1	0.29	0.45	0	1	0.69	0.46
HCC167	0	1	0.35	0.48	0	1	0.62	0.49
HCC168	0	1	0.00	0.01	0	0	0.00	0.00
HCC169	0	0	0.00	0.00	0	0	0.00	0.00
HCC170	0	1	0.00	0.01	0	0	0.00	0.00
HCC171	0	1	0.00	0.01	0	0	0.00	0.00
HCC172	0	1	0.00	0.01	0	0	0.00	0.00
HCC173	0	0	0.00	0.00	0	0	0.00	0.00

Variable	Datos 1998 (Todos los datos)				Datos 1998 (Datos de la clase)			
	Min	Max	Media	Std	Min	Max	Media	Std
HCC174	0	1	0.00	0.04	0	1	0.01	0.10
HCC175	0	1	0.00	0.04	0	1	0.02	0.12
HCC176	0	1	0.00	0.04	0	1	0.03	0.16
HCC177	0	1	0.00	0.04	0	1	0.03	0.16
HCC178	0	1	0.00	0.01	0	0	0.00	0.00
HCC179	0	1	0.05	0.21	0	1	0.15	0.36
HCC180	0	1	0.00	0.03	0	1	0.01	0.07
HCC181	0	1	0.00	0.06	0	1	0.04	0.20
HCC182	0	1	0.01	0.10	0	1	0.05	0.21
HCC183	0	1	0.34	0.48	0	1	0.43	0.50
HCC184	0	1	0.03	0.18	0	1	0.07	0.25
COMORB	0	55	6.95	4.93	0	42	16.92	8.74
RXTYPE	0	3	1.20	0.97	0	3	0.94	1.00
ACOVNY	0.00	1656008.10	7315.21	19884.29	121712.74	1656008.10	197388.53	129993.85

Tabla B.16 Principales estadísticas de los datos de 1998 (todos los datos y datos de la clase).

Variable	Datos 2000 (Todos los datos)			
	Min	Max	Media	Std
SEX	1	2	1.48	0.50
AGE	0	97	51.81	10.82
ELIG1	1	12	11.69	1.42
ELIG2	1	12	11.37	2.00
ACOVY	0.00	1653342.00	7355.13	19799.54
COVYI	0.00	1183460.00	2367.70	13588.18
COVYO	0.00	1653342.00	3417.42	10732.81
COVYD	0.00	77345.69	2012.69	2377.57
X_1	0.00	373.89	0.55	2.30
X_2	0.00	217.11	0.97	2.21
X_3	0.00	134.63	0.66	1.73
X_4	0.00	172.38	1.49	3.11
SLOPE	-63.07	45.64	0.25	1.13
OFFSET	-75.82	338.59	0.29	2.92
SUDDENNESS	-1.60	1.20	0.18	0.49
CONCAVITY	-2.00	2.00	-0.19	0.95
RMSE	-1.00	93.24	0.62	1.31
NLR	-1.00	1.67	0.59	0.34
COVYI_1	0.00	753314.00	548.26	5312.77
COVYI_2	0.00	815778.00	552.01	5366.36
COVYI_3	0.00	267767.00	381.98	4109.60
COVYI_4	0.00	1183460.00	885.45	8152.76
TLOS	0	0	0.00	0.00
NHOSP	0	0	0.00	0.00
HOS1	0	0	0.00	0.00
HOS2	0	0	0.00	0.00

Variable	Datos 2000 (Todos los datos)			
	Min	Max	Media	Std
HOS3	0	0	0.00	0.00
HOS4	0	0	0.00	0.00
COVYO_1	0.00	1653342.00	799.49	6121.68
COVYO_2	0.00	564919.00	855.11	3495.91
COVYO_3	0.00	139245.00	564.37	2118.07
COVYO_4	0.00	302376.00	1200.84	4097.72
COVYD_1	0.00	0.00	0.00	0.00
COVYD_2	0.00	21555.81	488.11	647.04
COVYD_3	0.00	25618.20	333.99	480.93
COVYD_4	0.00	26856.48	745.52	966.90
COSTRI4	-5	5	0.00	0.98
COSTRI43	-5	5	0.00	0.99
COSTRII4	-3	3	0.00	0.46
COSTRII43	-3	3	0.00	0.46
ACOVY_1	0.00	1653342.00	1347.75	8384.00
ACOVY_2	0.00	824466.08	1895.22	6916.52
ACOVY_3	0.00	275739.66	1280.34	4969.34
ACOVY_4	0.00	1238072.16	2831.82	10041.18
COSTDXCG	172.54	141735.05	7386.55	8538.44
SCORE	0.01	123.28	0.71	1.50
HCC001	0	1	0.00	0.04
HCC002	0	1	0.01	0.07
HCC003	0	1	0.00	0.05
HCC004	0	1	0.00	0.02
HCC005	0	1	0.00	0.03
HCC006	0	1	0.09	0.29
HCC007	0	1	0.00	0.07
HCC008	0	1	0.00	0.06
HCC009	0	1	0.01	0.08
HCC010	0	1	0.03	0.16
HCC011	0	1	0.00	0.04
HCC012	0	1	0.03	0.17
HCC013	0	1	0.03	0.18
HCC014	0	1	0.05	0.22
HCC015	0	1	0.02	0.14
HCC016	0	1	0.05	0.22
HCC017	0	1	0.03	0.17
HCC018	0	1	0.08	0.27
HCC019	0	1	0.67	0.47
HCC020	0	1	0.19	0.39
HCC021	0	1	0.00	0.04
HCC022	0	1	0.01	0.10
HCC023	0	1	0.02	0.15
HCC024	0	1	0.26	0.44

Variable	Datos 2000 (Todos los datos)			
	Min	Max	Media	Std
HCC025	0	1	0.00	0.04
HCC026	0	1	0.00	0.05
HCC027	0	1	0.00	0.04
HCC028	0	1	0.00	0.02
HCC029	0	1	0.01	0.09
HCC030	0	1	0.01	0.10
HCC031	0	1	0.01	0.07
HCC032	0	1	0.01	0.08
HCC033	0	1	0.00	0.06
HCC034	0	1	0.03	0.17
HCC035	0	1	0.00	0.03
HCC036	0	1	0.13	0.34
HCC037	0	1	0.01	0.09
HCC038	0	1	0.02	0.13
HCC039	0	1	0.04	0.20
HCC040	0	1	0.03	0.16
HCC041	0	1	0.02	0.13
HCC042	0	1	0.00	0.02
HCC043	0	1	0.25	0.43
HCC044	0	1	0.00	0.05
HCC045	0	1	0.00	0.05
HCC046	0	1	0.01	0.09
HCC047	0	1	0.04	0.19
HCC048	0	1	0.00	0.06
HCC049	0	1	0.00	0.05
HCC050	0	1	0.00	0.04
HCC051	0	1	0.00	0.03
HCC052	0	1	0.00	0.05
HCC053	0	1	0.00	0.06
HCC054	0	1	0.00	0.05
HCC055	0	1	0.03	0.17
HCC056	0	1	0.00	0.03
HCC057	0	1	0.00	0.02
HCC058	0	1	0.02	0.14
HCC059	0	1	0.01	0.08
HCC060	0	1	0.02	0.14
HCC061	0	1	0.00	0.00
HCC062	0	1	0.00	0.00
HCC063	0	1	0.00	0.00
HCC064	0	1	0.00	0.02
HCC065	0	1	0.00	0.02
HCC066	0	1	0.00	0.04
HCC067	0	1	0.00	0.02
HCC068	0	1	0.00	0.02

Variable	Datos 2000 (Todos los datos)			
	Min	Max	Media	Std
HCC069	0	1	0.00	0.04
HCC070	0	1	0.00	0.01
HCC071	0	1	0.02	0.15
HCC072	0	1	0.00	0.05
HCC073	0	1	0.00	0.03
HCC074	0	1	0.01	0.09
HCC075	0	1	0.00	0.03
HCC076	0	1	0.04	0.20
HCC077	0	1	0.00	0.02
HCC078	0	1	0.00	0.02
HCC079	0	1	0.01	0.10
HCC080	0	1	0.04	0.20
HCC081	0	1	0.01	0.09
HCC082	0	1	0.02	0.15
HCC083	0	1	0.02	0.14
HCC084	0	1	0.07	0.25
HCC085	0	1	0.00	0.05
HCC086	0	1	0.03	0.16
HCC087	0	1	0.00	0.01
HCC088	0	1	0.00	0.04
HCC089	0	1	0.00	0.06
HCC090	0	1	0.02	0.14
HCC091	0	1	0.32	0.47
HCC092	0	1	0.02	0.15
HCC093	0	1	0.02	0.14
HCC094	0	1	0.01	0.07
HCC095	0	1	0.00	0.04
HCC096	0	1	0.01	0.11
HCC097	0	1	0.02	0.12
HCC098	0	1	0.00	0.03
HCC099	0	1	0.00	0.02
HCC100	0	1	0.00	0.05
HCC101	0	1	0.00	0.03
HCC102	0	1	0.00	0.04
HCC103	0	1	0.00	0.04
HCC104	0	1	0.01	0.10
HCC105	0	1	0.03	0.16
HCC106	0	1	0.03	0.17
HCC107	0	1	0.00	0.02
HCC108	0	1	0.03	0.17
HCC109	0	1	0.01	0.08
HCC110	0	1	0.03	0.16
HCC111	0	1	0.00	0.04
HCC112	0	1	0.00	0.04

Variable	Datos 2000 (Todos los datos)			
	Min	Max	Media	Std
HCC113	0	1	0.02	0.15
HCC114	0	1	0.01	0.07
HCC115	0	1	0.08	0.27
HCC116	0	1	0.00	0.01
HCC117	0	1	0.00	0.04
HCC118	0	1	0.00	0.06
HCC119	0	1	0.03	0.16
HCC120	0	1	0.06	0.23
HCC121	0	1	0.01	0.09
HCC122	0	1	0.05	0.21
HCC123	0	1	0.05	0.22
HCC124	0	1	0.06	0.23
HCC125	0	1	0.00	0.06
HCC126	0	1	0.01	0.11
HCC127	0	1	0.25	0.43
HCC128	0	1	0.00	0.06
HCC129	0	0	0.00	0.00
HCC130	0	1	0.00	0.05
HCC131	0	1	0.01	0.12
HCC132	0	1	0.00	0.06
HCC133	0	1	0.02	0.15
HCC134	0	1	0.01	0.08
HCC135	0	1	0.05	0.23
HCC136	0	1	0.03	0.16
HCC137	0	1	0.00	0.06
HCC138	0	1	0.01	0.12
HCC139	0	1	0.08	0.28
HCC140	0	1	0.06	0.23
HCC141	0	1	0.00	0.01
HCC142	0	1	0.00	0.04
HCC143	0	1	0.00	0.04
HCC144	0	1	0.00	0.07
HCC145	0	1	0.00	0.05
HCC146	0	1	0.00	0.04
HCC147	0	1	0.00	0.04
HCC148	0	1	0.00	0.06
HCC149	0	1	0.01	0.12
HCC150	0	1	0.00	0.00
HCC151	0	1	0.00	0.02
HCC152	0	1	0.06	0.24
HCC153	0	1	0.14	0.35
HCC154	0	1	0.00	0.01
HCC155	0	1	0.00	0.04
HCC156	0	1	0.00	0.04

Variable	Datos 2000 (Todos los datos)			
	Min	Max	Media	Std
HCC157	0	1	0.00	0.04
HCC158	0	1	0.00	0.04
HCC159	0	1	0.01	0.08
HCC160	0	1	0.00	0.03
HCC161	0	1	0.00	0.04
HCC162	0	1	0.13	0.34
HCC163	0	1	0.02	0.14
HCC164	0	1	0.01	0.11
HCC165	0	1	0.01	0.09
HCC166	0	1	0.25	0.43
HCC167	0	1	0.18	0.39
HCC168	0	0	0.00	0.00
HCC169	0	1	0.00	0.00
HCC170	0	1	0.00	0.01
HCC171	0	1	0.00	0.00
HCC172	0	1	0.00	0.01
HCC173	0	0	0.00	0.00
HCC174	0	1	0.00	0.04
HCC175	0	1	0.00	0.03
HCC176	0	1	0.00	0.04
HCC177	0	1	0.00	0.03
HCC178	0	0	0.00	0.00
HCC179	0	1	0.06	0.24
HCC180	0	1	0.00	0.04
HCC181	0	1	0.00	0.06
HCC182	0	1	0.01	0.11
HCC183	0	1	0.31	0.46
HCC184	0	1	0.03	0.16
COMORB	0	31	5.33	3.41
RXTYPE	0	3	1.40	0.92

Tabla B.17 Principales estadísticas de los datos de 2000 (todos los datos).

Variable	Datos 1997				Datos 1998				Datos 2000			
	Min	Max	Media	Std	Min	Max	Media	Std	Min	Max	Media	Std
SEX	1	2	1.40	0.49	1	2	1.43	0.50	1	2	1.43	0.50
AGE	3	7	5.66	0.98	3	8	5.91	1.00	2	8	5.89	0.91
ELIG1	1	4	3.92	0.38	1	4	3.90	0.41	1	4	3.92	0.42
ELIG2	1	4	3.77	0.64	1	4	3.79	0.59	1	4	3.82	0.57
ACOVY	1	6	3.91	1.77	1	6	3.69	1.73	1	6	3.66	1.88
COVYI	2	6	3.46	1.53	2	6	3.37	1.44	2	6	3.34	1.47
COVYO	1	6	3.81	1.91	1	6	3.51	1.87	1	6	3.52	1.94
COVYD	1	6	2.43	1.64	1	6	2.22	1.48	1	6	2.14	1.53
X_1	1	6	2.10	1.42	1	6	1.99	1.37	1	6	1.87	1.30
X_2	1	6	2.19	1.42	1	6	2.12	1.43	1	6	1.90	1.27

Variable	Datos 1997				Datos 1998				Datos 2000			
	Min	Max	Media	Std	Min	Max	Media	Std	Min	Max	Media	Std
X_3	1	6	2.14	1.44	1	6	2.10	1.45	1	6	1.96	1.36
X_4	1	6	2.34	1.50	1	6	2.10	1.31	1	6	2.22	1.42
SLOPE	1	3	2.22	0.85	1	3	2.18	0.85	1	3	2.37	0.79
OFFSET	1	2	1.44	0.50	1	2	1.41	0.49	1	2	1.29	0.45
SUDDENNESS	1	3	2.10	0.34	1	3	2.03	0.30	1	3	2.06	0.31
CONCAVITY	1	3	1.80	0.57	1	3	1.86	0.58	1	3	1.84	0.59
RMSE	1	5	3.30	1.54	1	5	3.17	1.61	1	5	3.19	1.57
NLR	1	5	2.74	1.27	1	5	2.61	1.27	1	5	2.76	1.32
COVYI_1	4	6	4.34	0.69	4	6	4.24	0.58	4	6	4.20	0.53
COVYI_2	4	6	4.31	0.64	4	6	4.32	0.64	4	6	4.25	0.56
COVYI_3	4	6	4.26	0.60	4	6	4.24	0.57	4	6	4.24	0.57
COVYI_4	3	6	3.73	1.08	3	6	3.64	0.96	3	6	3.71	1.03
TLOS	0	7	2.06	2.33	0	7	2.00	2.26	0	0	0.00	0.00
NHOSP	0	3	1.21	1.09	0	3	1.25	1.13	0	0	0.00	0.00
HOS1	1	6	1.82	1.65	1	6	1.59	1.42	1	1	1.00	0.00
HOS2	1	6	1.81	1.62	1	6	1.84	1.63	1	1	1.00	0.00
HOS3	1	6	1.46	1.24	1	6	1.45	1.28	1	1	1.00	0.00
HOS4	1	6	2.02	1.80	1	6	1.85	1.62	1	1	1.00	0.00
COVYO_1	1	6	2.95	1.95	1	6	2.77	1.92	1	6	2.63	1.89
COVYO_2	1	6	3.18	1.96	1	6	3.11	1.93	1	6	2.89	1.91
COVYO_3	1	6	3.14	1.96	1	6	3.16	1.92	1	6	2.99	1.93
COVYO_4	1	6	3.57	1.96	1	6	3.28	1.92	1	6	3.43	1.99
COVYD_1	1	6	2.18	1.55	1	6	2.20	1.54	6	6	6.00	0.00
COVYD_2	1	6	2.30	1.59	1	6	2.16	1.49	1	6	2.05	1.45
COVYD_3	1	6	2.30	1.69	1	6	1.96	1.34	1	6	2.05	1.50
COVYD_4	1	6	2.39	1.64	1	6	2.11	1.40	1	6	2.06	1.48
COSTRI4	0	5	0.57	1.07	0	5	0.64	1.11	0	5	0.43	0.90
COSTRI43	0	5	0.37	0.87	0	5	0.47	0.96	0	5	0.36	0.82
COSTRII4	0	3	0.41	0.78	0	3	0.43	0.82	0	3	0.34	0.73
COSTRII43	0	3	0.32	0.68	0	3	0.34	0.74	0	3	0.29	0.66
ACOVY_1	1	6	3.10	1.96	1	6	2.94	1.83	1	6	2.68	1.87
ACOVY_2	1	6	3.34	1.85	1	6	3.23	1.88	1	6	3.00	1.89
ACOVY_3	1	6	3.25	1.86	1	6	3.22	1.79	1	6	3.07	1.88
ACOVY_4	1	6	3.65	1.85	1	6	3.39	1.81	1	6	3.57	1.91
COSTDXCG	1	6	3.92	1.77	1	6	3.86	1.64	1	6	3.72	1.79
SCORE	1	6	4.10	1.79	1	6	3.89	1.69	1	6	3.84	1.88
HCC001	0	0	0.00	0.00	0	1	0.01	0.07	0	1	0.00	0.05
HCC002	0	1	0.11	0.31	0	1	0.07	0.25	0	1	0.11	0.32
HCC003	0	1	0.01	0.08	0	1	0.01	0.10	0	1	0.02	0.15
HCC004	0	1	0.01	0.08	0	0	0.00	0.00	0	1	0.00	0.05
HCC005	0	1	0.02	0.14	0	1	0.02	0.12	0	1	0.02	0.15
HCC006	0	1	0.25	0.43	0	1	0.22	0.41	0	1	0.15	0.36
HCC007	0	1	0.06	0.24	0	1	0.04	0.20	0	1	0.06	0.23
HCC008	0	1	0.03	0.18	0	1	0.04	0.19	0	1	0.03	0.18

Variable	Datos 1997				Datos 1998				Datos 2000			
	Min	Max	Media	Std	Min	Max	Media	Std	Min	Max	Media	Std
HCC009	0	1	0.06	0.23	0	1	0.07	0.25	0	1	0.03	0.18
HCC010	0	1	0.12	0.33	0	1	0.07	0.26	0	1	0.05	0.22
HCC011	0	1	0.01	0.08	0	1	0.02	0.14	0	1	0.01	0.08
HCC012	0	1	0.07	0.25	0	1	0.09	0.28	0	1	0.04	0.19
HCC013	0	1	0.10	0.30	0	1	0.07	0.25	0	1	0.04	0.20
HCC014	0	1	0.08	0.27	0	1	0.07	0.26	0	1	0.02	0.15
HCC015	0	1	0.30	0.46	0	1	0.27	0.44	0	1	0.19	0.39
HCC016	0	1	0.30	0.46	0	1	0.28	0.45	0	1	0.13	0.33
HCC017	0	1	0.21	0.41	0	1	0.24	0.43	0	1	0.04	0.19
HCC018	0	1	0.27	0.44	0	1	0.32	0.47	0	1	0.14	0.35
HCC019	0	1	0.91	0.29	0	1	0.89	0.31	0	1	0.41	0.49
HCC020	0	1	0.50	0.50	0	1	0.52	0.50	0	1	0.39	0.49
HCC021	0	1	0.03	0.18	0	1	0.02	0.12	0	1	0.02	0.14
HCC022	0	1	0.08	0.27	0	1	0.07	0.25	0	1	0.05	0.22
HCC023	0	1	0.25	0.43	0	1	0.19	0.39	0	1	0.19	0.39
HCC024	0	1	0.37	0.48	0	1	0.38	0.49	0	1	0.21	0.41
HCC025	0	1	0.02	0.14	0	1	0.01	0.10	0	1	0.01	0.10
HCC026	0	1	0.03	0.18	0	1	0.03	0.17	0	1	0.01	0.11
HCC027	0	1	0.02	0.14	0	1	0.01	0.10	0	1	0.01	0.09
HCC028	0	1	0.01	0.08	0	1	0.02	0.12	0	1	0.00	0.05
HCC029	0	1	0.08	0.28	0	1	0.06	0.23	0	1	0.02	0.15
HCC030	0	1	0.06	0.23	0	1	0.05	0.22	0	1	0.04	0.19
HCC031	0	1	0.08	0.27	0	1	0.08	0.28	0	1	0.06	0.23
HCC032	0	1	0.05	0.22	0	1	0.03	0.17	0	1	0.03	0.17
HCC033	0	1	0.02	0.14	0	1	0.02	0.14	0	1	0.00	0.07
HCC034	0	1	0.12	0.33	0	1	0.11	0.31	0	1	0.09	0.28
HCC035	0	1	0.02	0.14	0	1	0.01	0.07	0	1	0.00	0.07
HCC036	0	1	0.50	0.50	0	1	0.41	0.49	0	1	0.22	0.42
HCC037	0	1	0.10	0.30	0	1	0.07	0.26	0	1	0.06	0.25
HCC038	0	1	0.08	0.27	0	1	0.03	0.16	0	1	0.04	0.20
HCC039	0	1	0.08	0.27	0	1	0.13	0.34	0	1	0.06	0.25
HCC040	0	1	0.01	0.08	0	1	0.04	0.19	0	1	0.02	0.14
HCC041	0	1	0.13	0.34	0	1	0.10	0.30	0	1	0.05	0.22
HCC042	0	0	0.00	0.00	0	0	0.00	0.00	0	1	0.00	0.07
HCC043	0	1	0.48	0.50	0	1	0.53	0.50	0	1	0.28	0.45
HCC044	0	1	0.06	0.24	0	1	0.04	0.19	0	1	0.04	0.19
HCC045	0	1	0.04	0.20	0	1	0.03	0.17	0	1	0.03	0.17
HCC046	0	1	0.11	0.31	0	1	0.07	0.25	0	1	0.05	0.21
HCC047	0	1	0.39	0.49	0	1	0.32	0.47	0	1	0.16	0.37
HCC048	0	1	0.03	0.16	0	1	0.04	0.20	0	1	0.04	0.19
HCC049	0	0	0.00	0.00	0	1	0.01	0.10	0	1	0.02	0.12
HCC050	0	1	0.03	0.18	0	1	0.04	0.19	0	1	0.01	0.09
HCC051	0	0	0.00	0.00	0	1	0.01	0.07	0	1	0.00	0.05
HCC052	0	1	0.01	0.08	0	1	0.01	0.07	0	1	0.01	0.09

Variable	Datos 1997				Datos 1998				Datos 2000			
	Min	Max	Media	Std	Min	Max	Media	Std	Min	Max	Media	Std
HCC053	0	1	0.01	0.08	0	1	0.01	0.10	0	1	0.01	0.09
HCC054	0	0	0.00	0.00	0	0	0.00	0.00	0	1	0.00	0.05
HCC055	0	1	0.05	0.22	0	1	0.12	0.32	0	1	0.06	0.25
HCC056	0	1	0.01	0.08	0	1	0.03	0.16	0	1	0.01	0.08
HCC057	0	1	0.01	0.08	0	1	0.01	0.10	0	0	0.00	0.00
HCC058	0	1	0.03	0.18	0	1	0.05	0.22	0	1	0.02	0.13
HCC059	0	1	0.02	0.14	0	1	0.04	0.20	0	1	0.00	0.07
HCC060	0	1	0.08	0.28	0	1	0.07	0.25	0	1	0.02	0.15
HCC061	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC062	0	1	0.01	0.08	0	0	0.00	0.00	0	0	0.00	0.00
HCC063	0	0	0.00	0.00	0	0	0.00	0.00	0	1	0.00	0.05
HCC064	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC065	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC066	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC067	0	1	0.01	0.08	0	0	0.00	0.00	0	0	0.00	0.00
HCC068	0	0	0.00	0.00	0	0	0.00	0.00	0	1	0.00	0.05
HCC069	0	1	0.01	0.08	0	1	0.01	0.10	0	1	0.01	0.08
HCC070	0	0	0.00	0.00	0	1	0.01	0.07	0	0	0.00	0.00
HCC071	0	1	0.13	0.34	0	1	0.17	0.38	0	1	0.11	0.31
HCC072	0	0	0.00	0.00	0	1	0.02	0.12	0	1	0.00	0.05
HCC073	0	0	0.00	0.00	0	0	0.00	0.00	0	1	0.00	0.05
HCC074	0	1	0.06	0.23	0	1	0.03	0.16	0	1	0.03	0.18
HCC075	0	1	0.01	0.12	0	1	0.01	0.07	0	1	0.01	0.08
HCC076	0	1	0.08	0.27	0	1	0.12	0.32	0	1	0.07	0.26
HCC077	0	0	0.00	0.00	0	0	0.00	0.00	0	1	0.00	0.07
HCC078	0	1	0.01	0.08	0	0	0.00	0.00	0	1	0.01	0.09
HCC079	0	1	0.09	0.29	0	1	0.10	0.30	0	1	0.10	0.30
HCC080	0	1	0.25	0.43	0	1	0.29	0.46	0	1	0.25	0.43
HCC081	0	1	0.03	0.18	0	1	0.07	0.26	0	1	0.05	0.22
HCC082	0	1	0.10	0.31	0	1	0.12	0.33	0	1	0.08	0.28
HCC083	0	1	0.12	0.32	0	1	0.11	0.32	0	1	0.04	0.20
HCC084	0	1	0.26	0.44	0	1	0.36	0.48	0	1	0.14	0.35
HCC085	0	1	0.03	0.16	0	1	0.02	0.12	0	1	0.04	0.20
HCC086	0	1	0.11	0.31	0	1	0.18	0.38	0	1	0.13	0.33
HCC087	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC088	0	1	0.03	0.16	0	1	0.03	0.16	0	1	0.01	0.10
HCC089	0	1	0.07	0.25	0	1	0.06	0.23	0	1	0.04	0.21
HCC090	0	1	0.03	0.18	0	1	0.03	0.16	0	1	0.02	0.15
HCC091	0	1	0.44	0.50	0	1	0.50	0.50	0	1	0.31	0.46
HCC092	0	1	0.06	0.24	0	1	0.12	0.33	0	1	0.10	0.30
HCC093	0	1	0.12	0.32	0	1	0.12	0.33	0	1	0.07	0.25
HCC094	0	1	0.18	0.38	0	1	0.13	0.34	0	1	0.02	0.13
HCC095	0	1	0.01	0.08	0	1	0.02	0.14	0	1	0.01	0.10
HCC096	0	1	0.07	0.25	0	1	0.11	0.31	0	1	0.07	0.25

Variable	Datos 1997				Datos 1998				Datos 2000			
	Min	Max	Media	Std	Min	Max	Media	Std	Min	Max	Media	Std
HCC097	0	1	0.08	0.28	0	1	0.10	0.30	0	1	0.04	0.19
HCC098	0	0	0.00	0.00	0	1	0.02	0.12	0	1	0.00	0.05
HCC099	0	1	0.01	0.12	0	1	0.01	0.07	0	1	0.00	0.07
HCC100	0	1	0.01	0.08	0	1	0.03	0.16	0	1	0.01	0.09
HCC101	0	0	0.00	0.00	0	0	0.00	0.00	0	1	0.01	0.09
HCC102	0	1	0.01	0.08	0	0	0.00	0.00	0	1	0.00	0.05
HCC103	0	1	0.01	0.12	0	1	0.02	0.12	0	0	0.00	0.00
HCC104	0	1	0.12	0.32	0	1	0.11	0.31	0	1	0.10	0.31
HCC105	0	1	0.21	0.41	0	1	0.26	0.44	0	1	0.14	0.35
HCC106	0	1	0.19	0.40	0	1	0.24	0.43	0	1	0.10	0.29
HCC107	0	1	0.02	0.14	0	1	0.01	0.07	0	1	0.00	0.07
HCC108	0	1	0.07	0.25	0	1	0.08	0.28	0	1	0.08	0.27
HCC109	0	1	0.10	0.30	0	1	0.07	0.26	0	1	0.03	0.16
HCC110	0	1	0.04	0.20	0	1	0.06	0.23	0	1	0.03	0.17
HCC111	0	1	0.03	0.18	0	1	0.03	0.16	0	1	0.02	0.12
HCC112	0	1	0.03	0.16	0	1	0.02	0.12	0	1	0.01	0.08
HCC113	0	1	0.24	0.43	0	1	0.20	0.40	0	1	0.09	0.29
HCC114	0	1	0.08	0.28	0	1	0.08	0.28	0	1	0.06	0.24
HCC115	0	1	0.25	0.43	0	1	0.22	0.42	0	1	0.09	0.29
HCC116	0	0	0.00	0.00	0	1	0.01	0.07	0	0	0.00	0.00
HCC117	0	1	0.01	0.08	0	0	0.00	0.00	0	1	0.01	0.09
HCC118	0	1	0.03	0.16	0	1	0.02	0.14	0	1	0.02	0.12
HCC119	0	1	0.18	0.38	0	1	0.20	0.40	0	1	0.14	0.35
HCC120	0	1	0.20	0.40	0	1	0.22	0.41	0	1	0.12	0.33
HCC121	0	1	0.04	0.20	0	1	0.04	0.20	0	1	0.01	0.08
HCC122	0	1	0.05	0.22	0	1	0.08	0.27	0	1	0.08	0.26
HCC123	0	1	0.10	0.31	0	1	0.13	0.34	0	1	0.11	0.31
HCC124	0	1	0.11	0.31	0	1	0.06	0.24	0	1	0.06	0.25
HCC125	0	1	0.01	0.08	0	0	0.00	0.00	0	1	0.02	0.13
HCC126	0	1	0.03	0.16	0	1	0.03	0.16	0	1	0.02	0.14
HCC127	0	1	0.37	0.49	0	1	0.26	0.44	0	1	0.21	0.41
HCC128	0	1	0.07	0.25	0	1	0.04	0.19	0	1	0.04	0.20
HCC129	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC130	0	1	0.12	0.32	0	1	0.13	0.34	0	1	0.14	0.35
HCC131	0	1	0.39	0.49	0	1	0.42	0.49	0	1	0.22	0.42
HCC132	0	1	0.07	0.25	0	1	0.10	0.30	0	1	0.01	0.08
HCC133	0	1	0.09	0.29	0	1	0.10	0.30	0	1	0.05	0.21
HCC134	0	1	0.02	0.14	0	1	0.01	0.10	0	1	0.01	0.09
HCC135	0	1	0.17	0.38	0	1	0.19	0.39	0	1	0.11	0.31
HCC136	0	1	0.29	0.46	0	1	0.34	0.48	0	1	0.04	0.20
HCC137	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC138	0	1	0.02	0.14	0	1	0.02	0.14	0	1	0.02	0.13
HCC139	0	1	0.08	0.28	0	1	0.06	0.24	0	1	0.05	0.22
HCC140	0	1	0.07	0.25	0	1	0.11	0.32	0	1	0.03	0.17

Variable	Datos 1997				Datos 1998				Datos 2000			
	Min	Max	Media	Std	Min	Max	Media	Std	Min	Max	Media	Std
HCC141	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC142	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC143	0	0	0.00	0.00	0	1	0.01	0.07	0	0	0.00	0.00
HCC144	0	0	0.00	0.00	0	1	0.01	0.10	0	0	0.00	0.00
HCC145	0	1	0.01	0.08	0	1	0.01	0.07	0	1	0.00	0.05
HCC146	0	0	0.00	0.00	0	1	0.01	0.07	0	0	0.00	0.00
HCC147	0	1	0.01	0.08	0	1	0.01	0.07	0	1	0.00	0.05
HCC148	0	1	0.03	0.18	0	1	0.06	0.24	0	1	0.04	0.20
HCC149	0	1	0.16	0.37	0	1	0.16	0.37	0	1	0.05	0.22
HCC150	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC151	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC152	0	1	0.23	0.42	0	1	0.21	0.41	0	1	0.18	0.39
HCC153	0	1	0.26	0.44	0	1	0.30	0.46	0	1	0.13	0.34
HCC154	0	1	0.01	0.08	0	0	0.00	0.00	0	0	0.00	0.00
HCC155	0	0	0.00	0.00	0	1	0.01	0.10	0	1	0.00	0.07
HCC156	0	1	0.01	0.08	0	0	0.00	0.00	0	1	0.01	0.08
HCC157	0	0	0.00	0.00	0	1	0.01	0.07	0	1	0.00	0.07
HCC158	0	1	0.02	0.14	0	1	0.01	0.10	0	1	0.01	0.10
HCC159	0	1	0.01	0.12	0	1	0.04	0.19	0	1	0.02	0.15
HCC160	0	1	0.01	0.08	0	0	0.00	0.00	0	1	0.00	0.05
HCC161	0	1	0.06	0.23	0	1	0.04	0.20	0	1	0.02	0.15
HCC162	0	1	0.23	0.42	0	1	0.30	0.46	0	1	0.19	0.40
HCC163	0	1	0.10	0.31	0	1	0.05	0.22	0	1	0.05	0.22
HCC164	0	1	0.23	0.43	0	1	0.16	0.37	0	1	0.20	0.40
HCC165	0	1	0.09	0.29	0	1	0.05	0.22	0	1	0.02	0.15
HCC166	0	1	0.67	0.47	0	1	0.69	0.46	0	1	0.58	0.49
HCC167	0	1	0.67	0.47	0	1	0.62	0.49	0	1	0.14	0.35
HCC168	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC169	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC170	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC171	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC172	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC173	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC174	0	1	0.03	0.18	0	1	0.01	0.10	0	1	0.02	0.15
HCC175	0	1	0.03	0.16	0	1	0.02	0.12	0	1	0.01	0.08
HCC176	0	1	0.03	0.16	0	1	0.03	0.16	0	1	0.02	0.13
HCC177	0	1	0.02	0.14	0	1	0.03	0.16	0	1	0.01	0.08
HCC178	0	0	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00
HCC179	0	1	0.23	0.43	0	1	0.15	0.36	0	1	0.19	0.39
HCC180	0	1	0.01	0.08	0	1	0.01	0.07	0	1	0.00	0.05
HCC181	0	1	0.03	0.16	0	1	0.04	0.20	0	1	0.06	0.23
HCC182	0	1	0.04	0.20	0	1	0.05	0.21	0	1	0.06	0.24
HCC183	0	1	0.44	0.50	0	1	0.43	0.50	0	1	0.37	0.48
HCC184	0	1	0.09	0.29	0	1	0.07	0.25	0	1	0.06	0.23

Variable	Datos 1997				Datos 1998				Datos 2000			
	Min	Max	Media	Std	Min	Max	Media	Std	Min	Max	Media	Std
COMORB	1	6	3.10	1.04	1	6	3.07	0.99	1	5	2.36	0.81
RXTYPE	0	3	0.90	0.97	0	3	0.94	1.00	0	3	1.22	1.01

Tabla B.18 Principales estadísticas de los datos con *coarse grain* de 1997, 1998 y 2000 (todos los datos).

Referencias

[1] ACM (*Association for Computing Machinery*). *ACM Special Interest Group on Knowledge Discovey and Data Mining*. Septiembre 2006. [En línea] Disponible: http://www.acm.org/sigs/sigkdd/charter.php

[2] G. F. Anderson y J. Knickman, *Patterns of Expenditures Among High Utilizers of Medical Care Services*, *Medical Care* 22 (2), 143-148. (1984).

[3] P. M. Atkinson y A. R. L Tatnall. *Neural networks in remote sensing*. International Journal of Remote Sensing, vol. 18, no. 4, 1997.

[4] M. L. Berk y A. C. Monheit. *The Concentration of Health Care Expenditures Revisited*, *Health Affairs* 20 (2), 9-18 (2001).

[5] S. Bow. *Pattern Recognition and Image Processing*. Marcel Dekker. 1992.

[6] P. Domingos and M. Pazzani. *Beyond independence: Conditions for the optimality of the simple Bayesian classifier.* Proceedings of the 13th International Conference on Machine Learning. Morgan Kaufmann. 1996.

[7] R. O. Duda, P. E. Hart y D. G. Stork. *Pattern Classification*. John Wiley & Sons. 1973.

[8] R. O. Duda, P. E. Hart y D. G. Stork. *Pattern Classification*. 2a ed. John Wiley & Sons. 2001.

[9] C. Elkan. *Magical Thinking in Data Mining: Lessons from CoIL Challenge 2000*. Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD'01).

[10] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth y R. Uthurusamy. *Advances in Knowledge Discover and Data Mining*. 1996. The MIT Press.

[11] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth. *The KDD Process for extracting Useful Knowledge from Volumes of Data. Communications of the ACM*. November 1996. Vol 39, No. 11.

[12] J. A. Freeman, D. M. Skapura. *Neural Networks*. Addison-Wesley. 1992.

[13] A. A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag. 2002.

[14] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. 1989.

[15] D. Hand, H. Mannila y P. Smith. *Principles of Data Mining*. MIT Press. 2001.

[16] J. Hernández, M. J. Ramírez y C. Ferri. *Introducción a la Minería de Datos*. Pearson Educación S. A. 2004.

[17] A. K. Jain. *Handbook of Pattern Recognition and Image Processing*. Academic Press. 1986.

[18] S. K. Jenson y F. A. Waltz. Principal component analysis and canonical analysis in remote sensing. *Proceedings of American Photogrammetric Soc. 45th Ann. Meeting*. 1979.

[19] D. E. Johnson. *Métodos Multivariados Aplicados al Análisis de Datos*. International Thomson Editores. 2000.

[20] W. Klösgen y J. M. Zytkow. *Handbook of Data Mining and Knowledge Discovery*. Oxford. 2001.

[21] R. P. Lippman. *An introduction to computing with neural nets*. IEEE ASSP Magazine, vol 2, 1987.

[22] B. Müller y J. Reinhardt. Neural Networks. Springer-Verlag. 1991.

[23] N. J. Nilsson. *Learning machines*. Morgan Kaufmann Publishers. 1990.

[24] S. K. Pal y P. Mitra. *Pattern Recognition Algorithms for Data Mining*. Chapman & HALL/CRC. 2004.

[25] G. Piatestky-Shapiro y W. Frawley, *Knowledge Discovery in Databases*. The AAAI Press. 1991.
[26] Springer Publisher. *Data Mining and Knowledge Discovery Journal*. Springer US. Septiembre 2006. [En línea] Disponible: http://springerlink.metapress.com/content/100254/
[27] C. R. Stephens y R. Sukumar. *An Introduction to Data Mining*. *The Handbook of Marketing Research*. SAGE Publications. 2006.
[28] C. R. Stephens, H. Waelbroeck, S. Talley y R. Cruz. *Predicting Healthcare Costs using a Classifier System*. GECCO 2004 (Poster Session).
[29] I. H. Witten y E. Frank. *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann. 2000.