PREMIO U.C.M. DE INVESTIGACIÓN 2006 LÍNEA 3000

Predicción de crisis empresariales en seguros no vida, mediante árboles de decisión y reglas de clasificación

Zuleyka Díaz Martínez

UCM EDITORIAL COMPLUTENSE

Predicción de crisis empresariales en seguros no vida mediante árboles de decisión y reglas de clasificación

Zuleyka Díaz Martínez

COLECCIÓN: **LÍNEA 3000**

UCM EDITORIAL COMPLUTENSE

Donoso Cortés, 63 - 4.ª planta. 28015 Madrid
Tels.: 91 394 64 60/1. Fax: 91 394 64 58
ecsa@rect.ucm.es
www.editorialcomplutense.com

Primera edición:
Octubre de 2007

Diseño de cubierta:
Beatriz Alonso

Fotocomposición:
MCF Textos, S. A.

ISBN:
978-84-7491-882-3

Índice

Introducción

Dentro de la variedad de problemas a los que el análisis financiero intenta hacer frente, el de la detección precoz de la insolvencia empresarial ha sido objeto de interés constante no sólo en el ámbito académico sino también por parte de un amplio abanico de usuarios relacionados con el mundo empresarial, debido al gran número de agentes e intereses afectados cuando una insolvencia tiene lugar.

El mencionado problema adquiere una mayor dimensión, si cabe, cuando se manifiesta en el sector del seguro, dada la importancia que éste supone para el conjunto de la actividad económica tanto por su creciente aportación en términos cuantitativos como por la relevancia de la labor que desempeña. Por ello, además de por la escasez relativa de trabajos existentes en nuestro país en relación al tema tratado, es por lo que nos planteamos la realización del presente trabajo.

Como es bien sabido, en la actualidad la industria aseguradora se encuentra inmersa en el ambicioso proyecto comunitario denominado *Solvencia II*, encaminado al logro, mediante la reforma de las reglas existentes en la Unión Europea en relación con la solvencia de las entidades aseguradoras, de un mayor ajuste a las circunstancias específicas de cada una de ellas de los requisitos en materia de solvencia a las que las mismas se ven sometidas por parte de las autoridades reguladoras.

Aspecto esencial para la consecución de este objetivo es el logro de un mejor aprovechamiento de la información financiero-contable suministrada por las entidades sometidas a supervisión que permita extraer de dicha información toda su potencialidad latente en cuanto a caracte-

rizar la situación específica de cada compañía, su grado de cobertura de los riesgos asumidos y su posibilidad de incurrir en una situación de insolvencia que le impida hacer frente a los compromisos adquiridos.

La utilización de nuevos y más ricos y sofisticados modelos analíticos para el tratamiento de la información suministrada por las entidades aseguradoras constituirá entonces un requisito inexcusable para alcanzar el objetivo mencionado. Es en este marco donde se inscribe el problema que a nosotros nos interesa de la estimación o valoración de la solvencia de una entidad aseguradora.

La determinación de la solvencia futura de una empresa puede ser entendida como un problema de clasificación: dada una información inicial o conjunto de atributos asociados a una empresa, se pretende tomar la decisión de asignar esa empresa a una clase concreta de entre varias posibles.

Aunque esta tarea puede ser llevada a cabo, y probablemente con notable éxito, por un experto humano, es de primordial interés la utilización de técnicas analíticas o algoritmos que permitan eliminar la subjetividad, y el coste, que supone la intervención de dicho experto. Es por ello por lo que estaremos interesados en automatizar de algún modo el proceso de inferencia y toma de decisiones, utilizando con este propósito algún sistema de clasificación.

Tradicionalmente los sistemas de clasificación se han implementado a través de la aplicación de técnicas estadísticas de carácter paramétrico tales como el Análisis Discriminante o los modelos de variable de respuesta cualitativa (Logit, Probit, etc.), que utilizan ratios financieros como variables explicativas. Dadas las peculiaridades del sector asegurador, la mayoría de estas investigaciones de tipo empírico acerca del estudio de las crisis empresariales se centran en otros sectores de la economía. No obstante, en el sector de seguros español, cabe destacar los trabajos realizados por López Herrera *et al.* (1994), Mora Enguídanos (1994), Martín Peña *et al.* (1999) y Sanchis Arellano *et al.* (2003), donde se pone de manifiesto la utilidad de estos métodos para valorar la situación financiera de este tipo de empresas.

Sin embargo, aunque los resultados obtenidos han sido satisfactorios, todas estas técnicas presentan el inconveniente de que parten de

hipótesis más o menos restrictivas acerca de las propiedades distribucionales de las variables explicativas que, especialmente en el caso de la información contable, no se suelen cumplir. Además, dada su complejidad, puede resultar difícil extraer conclusiones de sus resultados para un usuario poco familiarizado con la técnica.

En un intento de superar estas limitaciones recientemente se ha sugerido el empleo de técnicas procedentes del campo de la Inteligencia Artificial, debido a su carácter de métodos no paramétricos o de distribución libre, que no precisan por tanto de hipótesis preestablecidas sobre las variables de partida. Dentro de este tipo de técnicas, para el problema que nos ocupa son de gran utilidad las que se encuadran en el *Machine Learning* (Aprendizaje Automático), el área de la Inteligencia Artificial que se ocupa del desarrollo de algoritmos capaces de «aprender» un modelo a partir de ejemplos. Un representante típico de esta categoría son las redes neuronales, de las que se han desarrollado un buen número de aplicaciones en los más variados campos.

En este sentido, se han propuesto en los últimos años distintos enfoques para la predicción del fracaso empresarial en el campo del seguro en España basados en técnicas procedentes de las áreas del Aprendizaje Automático y la Inteligencia Artificial, como redes neuronales (Martínez de Lejarza Esparducer, 1999), *rough set* (Segovia Vargas, 2003), algoritmos genéticos y *support vector machines* (Segovia Vargas *et al.*, 2004) o programación genética (Salcedo Sanz *et al.*, 2005). Pero aunque todas estas técnicas salvan algunos inconvenientes de las técnicas estadísticas tradicionales, o bien requieren de un cierto nivel de conocimiento e implicación del decisorio a la hora de establecer ciertos parámetros necesarios para su aplicación, o bien son modelos de «caja negra» que no permiten valorar la importancia relativa de las variables explicativas y, aunque proporcionen buenos resultados en términos de error de clasificación, no permiten establecer un modelo de predicción de insolvencias de interpretación sencilla.

Por estas razones planteamos este trabajo, que se inscribe dentro de esta tendencia, cada vez más acusada, a utilizar para el análisis de problemas de naturaleza económica y empresarial técnicas procedentes de las áreas del Aprendizaje Automático y la Inteligencia Artificial. Pero,

a diferencia de los trabajos citados en el párrafo anterior, pretendemos desarrollar un modelo fácilmente interpretable a través de la aplicación de un método de implementación sencilla, el algoritmo de inducción de árboles de decisión y reglas de clasificación C4.5 (Quinlan, 1993). Concretamente, el propósito de este trabajo es comprobar el grado de aplicabilidad del algoritmo C4.5 a la valoración de la solvencia de las empresas de seguros, siendo el fin último de nuestra investigación desarrollar un conjunto de modelos sencillos basados en ratios financieros en forma de árboles y reglas de decisión que ayuden a pronosticar las insolvencias en el sector del seguro. En nuestra opinión, la utilización de estos modelos eficientes de predicción de insolvencias facilitaría grandemente la labor de supervisión de las empresas aseguradoras permitiendo que los recursos limitados de la inspección se dirigiesen hacia aquéllas preseleccionadas como potencialmente insolventes. Asimismo, compararemos los resultados alcanzados con los que se obtienen mediante la aplicación de Regresión Logística.

El resto del trabajo se estructura de la siguiente forma: en el capítulo 1 se exponen los fundamentos teóricos del algoritmo C4.5. El capítulo 2 se refiere a todos los aspectos relativos a la selección de datos y variables que intervienen en nuestro estudio empírico. En el capítulo 3 se presentan los resultados obtenidos con el algoritmo C4.5 y su comparación con la Regresión Logística. Finalmente, en el capítulo 4 exponemos nuestras conclusiones.

Capítulo 1: Sistemas de inducción de árboles de decisión y conjuntos de reglas. El algoritmo C4.5

1.1. INTRODUCCIÓN

La mayoría de aplicaciones de inteligencia artificial a tareas de importancia práctica se basan en la construcción de un modelo del conocimiento adquirido por un experto humano. Este enfoque, que alcanzó un gran impacto a comienzos de los años ochenta, se ilustra en numerosos estudios reflejados en el libro *The Rise of the Expert Company* (Feigenbaum *et al.*, 1988). En ocasiones, el problema al que se enfrenta un experto puede entenderse como un problema de clasificación: asignar elementos a categorías o clases determinadas por sus propiedades. Por ejemplo, en la referencia anterior se cita un sistema desarrollado por American Express para ser empleado como ayuda en las autorizaciones de crédito. Las propiedades para esta aplicación son detalles de la transacción propuesta y la historia de crédito del cliente particular, y las clases se corresponden con la aprobación o el rechazo de la transacción.

Un modelo de clasificación, restringiéndonos a modelos ejecutables —aquellos que pueden ser representados como programas de ordenador—, puede ser construido de dos maneras muy diferentes. Por un lado, el modelo podría obtenerse con la intervención del experto o expertos relevantes; de esta manera, se han desarrollado muchos sistemas basados en conocimiento (sistemas expertos), a pesar de las bien conocidas limitaciones de este enfoque (Michie, 1987, 1989). Alternativamente, podrían examinarse clasificaciones almacenadas en una base

de datos y construir un modelo inductivamente, mediante la generalización a partir de ejemplos específicos (Quinlan, 1993).

El uso generalizado de los ordenadores en la sociedad de hoy implica que grandes cantidades de datos sean almacenadas electrónicamente. Estos datos se relacionan virtualmente con todas las facetas de la vida moderna y son un valioso recurso en la medida en que se disponga de las herramientas adecuadas para utilizarlos. Como ya hemos mencionado, los algoritmos de Aprendizaje Automático son un conjunto de técnicas que automáticamente construyen modelos que describen la estructura subyacente en un conjunto de datos, es decir, inducen un modelo u *output* a partir de un conjunto de observaciones o *input*. En la jerga del área, las observaciones individuales se denominan «instancias», «objetos» o «casos», y un conjunto de observaciones se denomina «conjunto de datos». Cada instancia contiene una serie de valores que miden ciertas propiedades de la misma. Estas propiedades se denominan «atributos». Los atributos pueden ser nominales, por ejemplo, colores, o numéricos, tanto enteros como reales. El espacio de todas las posibles combinaciones de los valores de los atributos se denomina «espacio de instancias», o también «espacio de características o de atributos».

Los modelos construidos tienen dos importantes aplicaciones. En primer lugar, en la medida en que representen de forma precisa la estructura subyacente en los datos, podrán ser utilizados para predecir propiedades de futuros elementos. En segundo lugar, en la medida en que resuman la información esencial de manera comprensible para el humano, podrán ser empleados para analizar el dominio del que proceden los datos.

Estas dos aplicaciones no son mutuamente excluyentes. Para que sea útil para el análisis, un modelo debe ser una representación precisa del dominio, lo que también le confiere utilidad para la predicción. Sin embargo, la reciprocidad no es necesariamente cierta: algunos modelos son diseñados exclusivamente para la predicción y su habilidad en esta faceta no implica su utilidad para el análisis. En muchas aplicaciones de minería de datos este enfoque de «caja negra» es un serio inconveniente porque los usuarios no pueden determinar cómo se deriva una predicción y asociar esta información con su conocimiento del domi-

nio. Esto imposibilita el uso de dichos modelos en aplicaciones críticas en las que un experto del dominio debe ser capaz de verificar el proceso de decisión que conduce a una predicción —por ejemplo, en aplicaciones de medicina—.

Este capítulo se centra en métodos para el aprendizaje de «modelos comprensibles». Con el término *modelo* se está indicando que estos métodos construyen un modelo o representación de la regularidad existente en los datos. Con el término *comprensibles* se está haciendo referencia al hecho de que estos modelos se pueden expresar de manera simbólica, en forma de conjunto de condiciones (a diferencia de otros métodos como las redes neuronales o las máquinas de vectores soporte) y, por tanto, resultan modelos inteligibles para los seres humanos.

Los árboles de decisión y los sistemas de reglas son poderosos predictores que incluyen una representación explícita del conocimiento inducido a partir de un conjunto de datos. Además, en comparación con otros modelos sofisticados, pueden ser generados muy rápidamente.

Dado un árbol de decisión o un conjunto de reglas, un usuario puede determinar manualmente cómo se deriva una predicción particular, y qué atributos son relevantes en la misma. Esto les convierte en herramientas extremadamente útiles en muchas aplicaciones de minería de datos donde son importantes tanto la capacidad predictiva como la utilidad para el análisis, esto es, la predicción y la explicación.

1.2. ÁRBOLES DE DECISIÓN

Los árboles de decisión son un modo de representación de la regularidad subyacente en los datos en forma de un conjunto de condiciones organizadas en una estructura jerárquica. Un árbol de decisión puede utilizarse para la exploración de datos con uno o varios de los siguientes fines (Murthy, 1998):

- *Descripción:* reducir una gran cantidad de datos transformándolos en una forma más compacta que preserva las características esenciales y proporciona un resumen preciso.

- *Clasificación:* descubrir si los datos contienen clases de objetos bien diferenciadas que puedan ser interpretadas de manera significativa en el contexto de una teoría sustantiva.
- *Generalización:* descubrir una relación entre variables independientes y dependientes que sea útil para predecir el valor de la variable dependiente en el futuro.

Los árboles de decisión automáticamente construidos a partir de un conjunto de datos han sido utilizados con éxito en muchas situaciones del mundo real. Su efectividad ha sido ampliamente comparada con la de otros métodos de exploración de datos y la de los expertos humanos.

En cuanto a la tarea de clasificación, varias ventajas de la clasificación basada en árboles de decisión han sido señaladas, entre ellas (Murthy, 1998):

– La adquisición de conocimiento a partir de ejemplos preclasificados salva el obstáculo que representa el adquirir el conocimiento mediante la intervención de un experto humano en el dominio.
– Los métodos basados en árboles son exploratorios en lugar de inferenciales. También son no paramétricos. Al no realizarse hipótesis sobre el modelo y la distribución de los datos, los árboles pueden modelizar un amplio rango de distribuciones de datos.
– La descomposición jerárquica supone un mejor uso de las características disponibles y una mayor eficiencia computacional en la clasificación.
– Al contrario que algunos métodos estadísticos, los árboles de clasificación pueden trabajar de la misma manera con datos unimodales y multimodales.
– Los árboles pueden usarse con la misma facilidad tanto en problemas deterministas como en problemas incompletos. (En dominios deterministas, la variable dependiente puede ser determinada perfectamente a partir de las variables independientes, mientras que en problemas incompletos, no.)

– Los árboles llevan a cabo la clasificación a través de una secuencia de preguntas simples, fáciles de entender, cuya semántica es intuitivamente clara para los expertos en el dominio. El formalismo de los árboles de decisión es en sí mismo muy intuitivo y atractivo para el entendimiento.

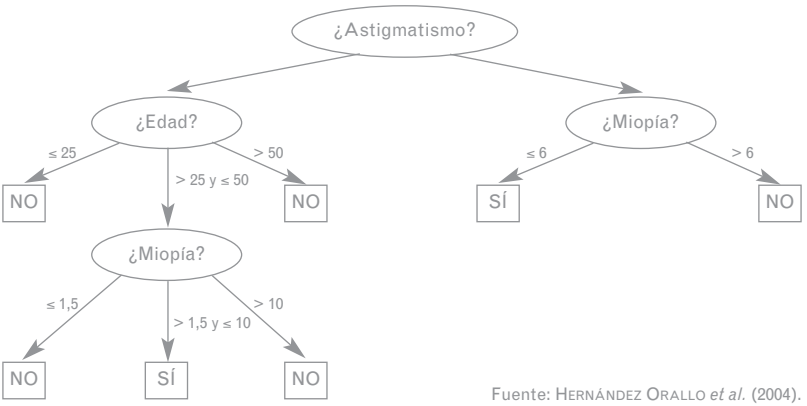
Por estas y otras razones, la metodología de árboles de decisión puede proporcionar una importante herramienta en la investigación y práctica en minería de datos. De hecho, muchas herramientas existentes en esta área se basan en la construcción de árboles de decisión a partir de conjuntos de datos.

Un árbol de decisión es una estructura en forma de árbol formada por nodos internos y externos conectados por ramas. Un nodo interno contiene una pregunta, es una unidad que evalúa una función de decisión para determinar cuál es el próximo nodo hijo a visitar. En contraste, un nodo externo, también llamado nodo hoja o nodo terminal, no tiene nodos hijos y se asocia con una etiqueta o valor que caracteriza a los datos que llegan al mismo.

En general, un árbol de decisión se emplea de la siguiente manera: en primer lugar, se presenta una instancia, un vector compuesto por varios atributos, al nodo inicial (o nodo raíz) del árbol de decisión. Dependiendo del resultado de la función de decisión usada por el nodo interno, el árbol nos conducirá hacia uno de los nodos hijos. Esto se repite hasta que se alcanza un nodo terminal y se asigna una etiqueta o valor a los datos de entrada.

Consideremos, por ejemplo, un hospital en el que se realizan operaciones de cirugía refractiva a personas miopes (Hernández Orallo *et al*., 2004). Obviamente, tales operaciones no están indicadas en muchos casos, y algunos podrían ser descartados en una primera fase para evitar riesgos potenciales o efectos secundarios. Aunque la indicación de dicha operación requiere un examen minucioso por parte del servicio de oftalmología del hospital, hay algunas condiciones que permiten determinar si, en principio, una persona está indicada para el estudio detallado. La siguiente figura (figura 1.1) muestra un ejemplo del árbol de decisión que se utiliza para admitir las solicitudes.

Figura 1.1. Árbol de decisión para determinar la recomendación de cirugía ocular



Fuente: Hernández Orallo *et al.* (2004).

Para saber si a un nuevo paciente se le ha de recomendar o no el estudio detallado, bastaría aplicar el árbol de decisión de la figura, realizando las preguntas de los nodos internos y siguiendo las respuestas hasta alguna de las hojas del árbol, catalogadas con un «sí» o un «no».

Una de las grandes ventajas de los árboles de decisión es que las opciones posibles a partir de una determinada condición son excluyentes, lo que permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar.

El árbol de decisión de la figura en concreto funciona como un clasificador, esto es, dado un nuevo individuo lo clasificará en una de las dos clases posibles: «sí» o «no».

Los árboles de decisión que se usan para problemas de clasificación son denominados a menudo «árboles de clasificación», y cada nodo terminal contiene una etiqueta que indica la clase predicha de un vector de características dado. Los árboles de decisión utilizados para problemas de regresión se denominan frecuentemente «árboles de regresión», y las etiquetas de los nodos terminales deben ser constantes o ecuaciones que especifican el valor *output* predicho de un vector *input* dado. En este caso, se suele distinguir entre «árboles de regresión» y «árboles de modelo», según que los nodos hoja contengan, respectivamente, constantes o funciones lineales multivariables.

Un árbol de decisión se denomina «binario» cuando cada nodo interno tiene exactamente dos hijos. Éstos son los más usados, debido a su simplicidad, aunque tampoco son infrecuentes los árboles que exhiben nodos con más de dos hijos. El árbol de la figura anterior se denomina «univariado», porque sólo un atributo se ve involucrado en la decisión adoptada en cada nodo. En cambio, los nodos de los árboles de decisión «multivariados» contienen preguntas sobre relaciones que envuelven más de un atributo, por ejemplo, combinaciones lineales de atributos (Brodley y Utgoff, 1995). Esto les convierte en predictores potencialmente más poderosos, sin embargo, también son más difíciles de interpretar y de generación más costosa computacionalmente.

Pasaremos ahora a estudiar cómo se construye un árbol de decisión a partir de un conjunto de datos, dejando de lado, al menos por el momento, las cuestiones relativas al uso e interpretación de un árbol ya construido.

La construcción automática de árboles de decisión ha sido virtualmente objeto de atención en todas las disciplinas en las que se han desarrollado métodos de exploración de datos. Tradicionalmente, ha sido considerada en los campos de la estadística y la ingeniería (reconocimiento de patrones). Más recientemente, la investigación en el campo de la Inteligencia Artificial, concretamente, en el área del Aprendizaje Automático, ha renovado el interés por el tema.

El trabajo sobre la inducción de árboles de decisión en estadística comienza debido a la necesidad de explorar datos de encuestas. Programas estadísticos como AID (Morgan y Sonquist, 1963), THAID (Morgan y Messenger, 1973) y CHAID (Kass, 1980) construyen árboles con el objetivo de descubrir las interacciones entre variables predictoras y dependientes.

La investigación sobre árboles de decisión en reconocimiento de patrones fue motivada por la necesidad de interpretar imágenes procedentes de sensores ubicados en satélites artificiales (Swain y Hauska, 1977).

El estudio de los árboles de decisión en particular y de los métodos de inducción en general surge en el aprendizaje automático para evitar los inconvenientes asociados a la adquisición del conocimiento (Feigenbaum, 1981) en los sistemas expertos.

Una amplia revisión multidisciplinar acerca de la construcción automática de árboles de decisión a partir de datos puede encontrarse en Murthy (1998).

Aquí nos centraremos en árboles de clasificación, puesto que la determinación de la solvencia futura de una empresa puede ser entendida como un problema de clasificación: dada una información inicial o conjunto de atributos asociados a una instancia, se pretende tomar la decisión de asignar a esa instancia una clase concreta de entre varias posibles. En otras palabras, dado un conjunto de ratios financieros que caracterizan a una empresa, pretendemos clasificarla como sana o fracasada.

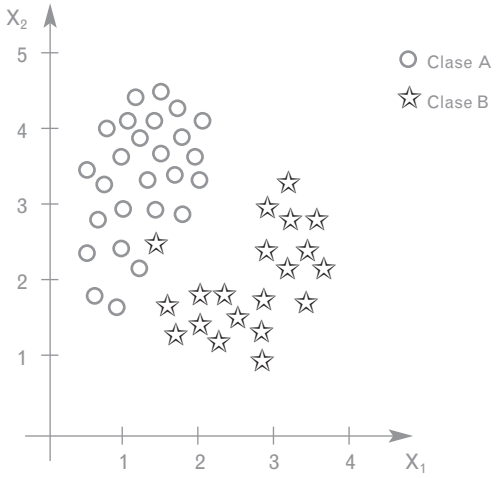
La estructura de condición y ramificación de un árbol de decisión es idónea para el problema que nos ocupa, el de clasificación. Debido al hecho de que la clasificación trata con clases o etiquetas disjuntas, es decir, una instancia es de una clase o de otra, pero no de varias clases a la vez, un árbol de decisión conducirá un ejemplo hasta una y sólo una hoja, asignándole, por tanto, una única clase. Para ello, las particiones, las condiciones existentes en el árbol, deben ser también disjuntas. Así, en el ejemplo de la figura anterior, una persona o tiene miopía > 6 o tiene miopía ≤ 6, pero no ambas cosas a la vez. Asimismo, dicha propiedad es exhaustiva, esto es, una de las dos condiciones se ha de cumplir.

Esto dio lugar al esquema básico de los principales algoritmos de aprendizaje de árboles de decisión, que construyen dichos árboles recursivamente siguiendo una estrategia descendente: el espacio de instancias se va partiendo de arriba abajo, empleando en cada partición una sola variable o atributo (una partición será entonces un conjunto de condiciones excluyentes y exhaustivas acerca de dicho atributo). Por ello se emplea el acrónimo TDIDT *(Top-Down Induction on Decision Trees)* para hacer referencia a la familia de algoritmos de construcción de árboles de decisión. La inducción *top-down* de árboles de decisión procede de acuerdo con la estrategia denominada *divide-and-conquer* ('divide y vencerás'), donde el conjunto de entrenamiento (el conjunto de datos usados para construir el árbol) se va partiendo en subconjuntos y el algoritmo se aplica recursivamente a cada uno de ellos.

Esto supone un enfoque radicalmente diferente al que siguen las técnicas estadísticas tradicionales o las redes neuronales al abordar el problema de la clasificación. Con éstas se trata de delimitar dos o más regiones de decisión insertando hipersuperficies más o menos complejas en el espacio *n*-dimensional que forman las *n* variables explicativas. En cambio, mediante la aplicación de algoritmos de inducción de árboles las regiones de decisión vienen definidas a través de una serie de hiperrectángulos.

Veamos un ejemplo sencillo de cómo operan estas técnicas. Supongamos que tenemos 46 instancias, 26 de ellas de la clase A y 20 de la clase B, que están caracterizadas por sólo dos variables explicativas, X_1 y X_2. En la siguiente figura (figura 1.2) se representa el problema.

Figura 1.2. Representación gráfica de un problema de clasificación



En primer lugar, seleccionaremos una de las variables y un punto de corte para la misma para configurar el nodo raíz del árbol. Más adelante veremos cómo se realiza esta selección. Ahora simplemente consideremos que la variable elegida es X_1 y el punto de corte 2.5, con lo que tendríamos la primera partición del conjunto de instancias (figura 1.3) que conforma el nodo raíz y los dos nodos hijos del árbol que se representa en la figura 1.4.

Figura 1.3. Primera partición del conjunto de instancias

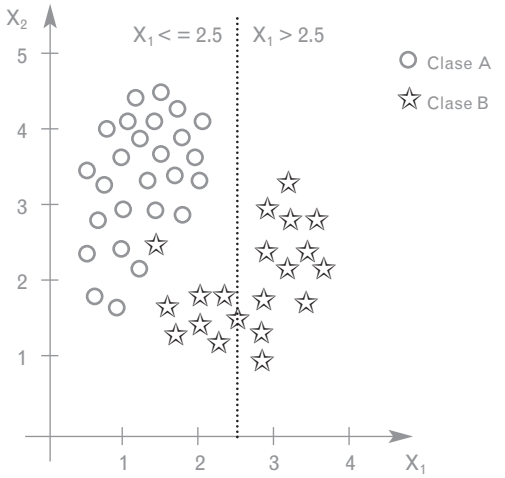
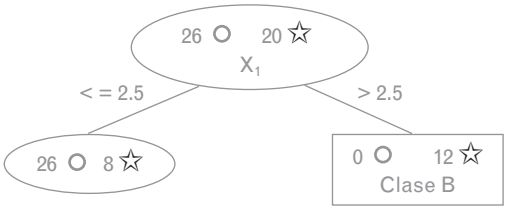


Figura 1.4. Árbol de clasificación (primera partición)



Hemos dividido el conjunto inicial en dos subconjuntos. Todos los puntos que verifican la condición $X_1 > 2.5$ son de la clase B, y formarían por tanto un nodo hoja. En cambio, la condición $X_1 \leq 2.5$ es verificada por puntos de las dos clases A y B, y no podríamos alcanzar una decisión en cuanto a la clasificación de los puntos que verifiquen esa condición. Tendríamos que considerar una nueva partición del subconjunto, y supongamos ahora que la variable seleccionada es X_2 y el punto de corte es 2 (figura 1.5), con lo que añadiremos dos nuevos nodos al árbol (figura 1.6).

Figura 1.5. Segunda partición del conjunto de instancias

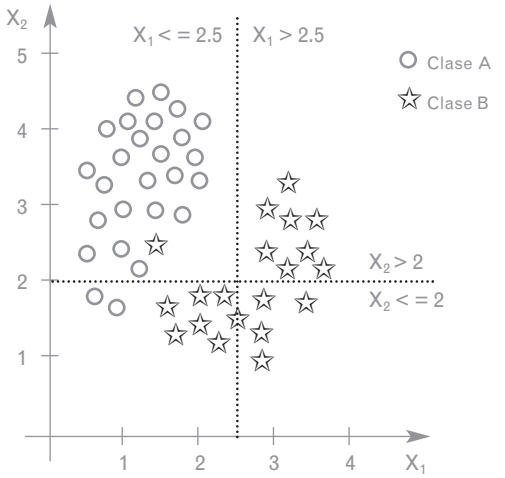
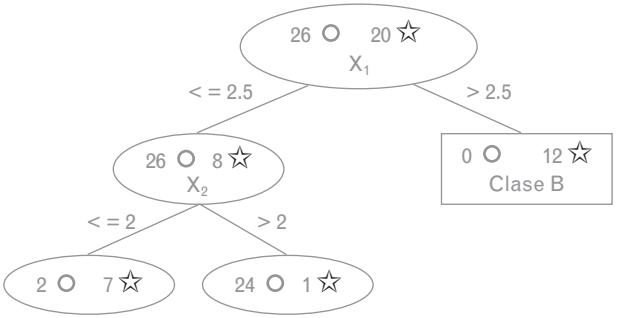
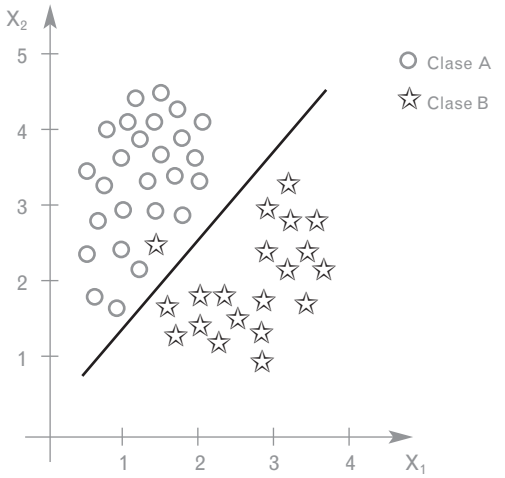


Figura 1.6. Árbol de clasificación (segunda partición)



De esta forma recursiva, insertando rectángulos en nuestro espacio bidimensional, vamos construyendo el árbol de decisión. Por contra, las técnicas tradicionales definirían una superficie de separación (en nuestro ejemplo, por trabajar sólo en dos dimensiones, una línea). Así, por ejemplo, a través de un análisis discriminante lineal las dos regiones de decisión vendrían delimitadas por la recta de la siguiente figura (figura 1.7).

Figura 1.7. Recta clasificadora



Los métodos de inducción de árboles de decisión mediante particiones recursivas del conjunto de casos de entrenamiento, como ya hemos mencionado, han sido un tópico activo en la investigación en inteligencia artificial (particularmente en aprendizaje automático) y en estadística (en particular, en análisis multivariante). En la literatura estadística destaca el trabajo de Friedman (1977), que sirvió como base para la construcción del sistema CART *(Classification and Regression Trees)*, descrito en Breiman *et al*. (1984). Éste es el más popular y utilizado de los algoritmos de inducción de árboles de decisión desarrollados por estadísticos.

Más o menos al mismo tiempo los sistemas de inducción de árboles de decisión comienzan a abordarse por un destacado investigador del área del aprendizaje automático, J. Ross Quinlan. Los algoritmos ID3 y C4.5 desarrollados por este autor a lo largo de las dos últimas décadas (Quinlan, 1979, 1983, 1986a, 1986b, 1987a, 1987b, 1988, 1989, 1990, 1991, 1993, 1996; Quinlan *et al*., 1987; Quinlan y Rivest, 1989) son posiblemente los más elaborados y los que, sin duda, han alcanzado una mayor repercusión.

Lo que diferencia fundamentalmente entre sí a los distintos algoritmos de «partición» existentes hasta la fecha es el criterio de selección

de particiones. Como ya hemos mencionado, la estructura de árbol se genera partiendo el conjunto de entrenamiento en subconjuntos más y más pequeños de una manera recursiva *top-down*. Comenzando con el conjunto de entrenamiento completo en el nodo raíz, se escoge una partición y se divide el conjunto en subconjuntos de acuerdo con ella. Recursivamente se va partiendo cada uno de los subconjuntos obtenidos hasta que todos ellos son «puros» o hasta que su pureza ya no pueda incrementarse. Un subconjunto es puro si las instancias que contiene son de una misma clase. El objetivo es alcanzar el máximo grado de pureza usando el menor número de particiones posible de manera que el árbol resultante sea pequeño y el número de instancias de cada subconjunto, grande.

La mayoría de métodos de construcción de árboles de decisión son algoritmos «voraces» *(greedy)*, en el sentido de que, en cada etapa, escogen la que consideran mejor opción en ese momento y no vuelven atrás. Es decir, una vez elegida la partición continúan hacia abajo la construcción del árbol y no vuelven a plantearse las particiones ya construidas. Por ello resulta esencial un criterio de selección de particiones que permita realizar una buena elección de la partición más prometedora sin demasiado esfuerzo computacional, ya que una mala elección de la partición, especialmente en las partes superiores del árbol, tendrá como consecuencia una degradación muy acusada en la calidad del mismo.

Durante las dos últimas décadas se han presentado diversos criterios de partición, tales como el «índice de Gini» empleado en CART (Breiman *et al*., 1984), la «ganancia de información» (Quinlan, 1986b) o el «ratio de ganancia» (Quinlan, 1993), entre otros. Todos ellos se basan en la idea de utilizar medidas de la pureza de un nodo para buscar particiones de máxima discriminación, es decir, que den lugar a los nodos más puros que sea posible, de manera que el algoritmo de aprendizaje vaya seleccionando las particiones que correspondan al mejor valor del criterio de partición empleado.

Como ya hemos señalado, un nodo puro será aquel al que sólo correspondan casos pertenecientes a una de las clases del problema. La bondad de una partición vendrá indicada por el decrecimiento de im-

pureza que se consigue con ella. La maximización de la bondad de una partición equivale, por tanto, a la minimización de la impureza del árbol generado por dicha partición.

Una función de impureza ϕ mide la impureza de un nodo. Dado un problema de clasificación con J clases diferentes, la función de impureza se define sobre el conjunto de las J-tuplas $(p_1, p_2, \ldots, p_j)$, donde los argumentos p_j, $j = 1, \ldots, J$, indican la probabilidad de que un caso pertenezca a la clase j en el subárbol actual. Lógicamente, $p_j \geq 0$ y $\sum_{j=1}^{J} p_j = 1$. Una función de impureza ϕ suele definirse como no negativa y poseer las siguientes propiedades (Breiman *et al*., 1984):

– Es máxima cuando la distribución de las clases en el nodo es uniforme, es decir, cuando el número de ejemplos correspondientes a cada una de las clases del problema es el mismo para todas ellas:

$$\phi(1/J, 1/J, \ldots, 1/J) = \text{máximo}.$$

– Es mínima e igual a cero cuando todos los casos pertenecen a la misma clase, es decir, la función ϕ alcanza sus J mínimos en $\phi(1, 0, \ldots, 0)$, $\phi(0, 1, \ldots, 0)$, …, $\phi(0, 0, \ldots, 1)$, y $\phi(1, 0, \ldots, 0) = \phi(0, 1, \ldots, 0) = \ldots = \phi(0, 0, \ldots, 1) = 0$.

– Es simétrica respecto a $p_1, p_2, \ldots, p_j$.

La impureza de un nodo t será entonces $I(t) = \phi(p_1, p_2, \ldots, p_j)$, donde p_j es el porcentaje de casos en el nodo t que pertenecen a la clase j. Análogamente, la impureza de un árbol T puede expresarse como:

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

donde $\tilde{T}$ es el conjunto de nodos terminales (nodos hoja) en el árbol T y $p(t)$ es la probabilidad de que un ejemplo dado corresponda a la hoja t.

Las funciones de impureza más utilizadas para un árbol de clasificación con J clases son el índice de Gini y la función de entropía:

Índice de Gini: $\phi_g(p_1, \ldots, p_j) = \sum_{\substack{i,j=1 \\ i \neq j}}^{J} p_i p_j = 1 - \sum_{j=1}^{J} p_j^2$.

Función de entropía: $\phi_e(p_1, \ldots, p_j) = -\sum_{j=1}^{J} p_j \log p_j$.

El índice de Gini es una medida de la diversidad de clases en un nodo del árbol. La diversidad será mínima cuando sólo haya una clase, $\phi_g(1, 0, \ldots, 0) = \phi_g(0, 1, \ldots, 0) = \phi_g(0, 0, \ldots, 1) = 0$ y será máxima cuando todas las clases sean equiprobables,

$$\phi_g\left(\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J}\right) = 1 - \sum_{j=1}^{J}\left(\frac{1}{J}\right)^2 = 1 - J\,\frac{1}{J^2} = 1 - \frac{1}{J} = \frac{J-1}{J}\,(\text{máximo}).$$

Es una medida de impureza utilizada en distintos algoritmos de construcción de árboles de decisión. En concreto, es la medida que se utiliza en CART (Breiman *et al*., 1984).

La entropía es también una medida de impureza muy utilizada y, en particular, es en la que se basan los algoritmos de Quinlan. Según los experimentos realizados durante los últimos años, parece ser que el criterio empleado por el algoritmo C4.5 manifiesta un comportamiento ligeramente superior al de Gini.

1.2.1. El algoritmo C4.5

Como ya hemos mencionado, el algoritmo C4.5 es el más popular de entre todos los algoritmos de aprendizaje de árboles de clasificación a partir de un conjunto de datos de ejemplo. Fue desarrollado por J. Ross Quinlan en la década de los años ochenta y principios de los noventa (Quinlan, 1993) como descendiente de un primer programa clasificador que fue denominado ID3 (Quinlan, 1979, 1983, 1986b).

La idea original se remonta al trabajo de Hoveland y Hunt de finales de los años cincuenta, que culminó en el libro pionero *Experiments in Induction* (Hunt *et al*., 1966), el cual describe experimentos con varias implementaciones de *Concept Learning Systems* (CLS).

El método de Hunt para construir un árbol de decisión a partir de un conjunto T de casos de entrenamiento es muy simple. Siendo las clases $\{C_1, C_2, \ldots, C_k\}$, hay tres posibilidades:

- Si T contiene uno o más casos y todos ellos son de la misma clase C_j, el árbol de decisión para el conjunto T será una hoja etiquetada con la clase C_j.

- Si T no contiene casos, es decir, si el conjunto de casos de entrenamiento queda vacío, el árbol de decisión será de nuevo una hoja, pero la clase asociada con dicha hoja tendrá que ser determinada utilizando información adicional. Por ejemplo, la etiqueta de la hoja podría ser escogida de acuerdo con algún conocimiento del dominio, como la clase mayoritaria en el conjunto. C4.5 usa la clase más frecuente en el padre de este nodo.
- Si T contiene casos de distintas clases, la idea es dividir T en subconjuntos que sean o conduzcan a agrupaciones uniformes de casos, esto es, conjuntos de casos correspondientes a una misma clase. Utilizando los casos de entrenamiento disponibles, se escoge una pregunta para ramificar el árbol. Dicha pregunta, basada en uno de los atributos predictivos, tendrá una o más respuestas mutuamente excluyentes $\{O_1, O_2, \ldots, O_n\}$. Entonces T se divide en los subconjuntos $T_1, T_2, \ldots, T_n$, donde T_i contiene todos los casos de T con respuesta O_i a la pregunta escogida. El árbol de decisión para T consistirá entonces en un nodo que identifica la pregunta realizada del que cuelgan tantas ramas como respuestas alternativas. El mismo método se aplica recursivamente para construir los subárboles correspondientes a cada hijo del nodo, teniendo en cuenta que a cada rama i, a cada hijo del nodo anterior, se le asigna el subconjunto de casos de entrenamiento T_i correspondientes a la respuesta O_i.

La sucesiva división del conjunto de casos de entrenamiento prosigue hasta que todos los subconjuntos están formados por casos pertenecientes a la misma clase, o hasta que ya no se encuentre ninguna pregunta para seguir ramificando el árbol que suponga alguna mejora.

Ilustremos el proceso con un ejemplo. Supongamos el pequeño conjunto de entrenamiento presentado en la siguiente tabla (tabla 1.1), donde cada fila representa un caso y la clase asociada al mismo se corresponde con la recomendación de jugar o no a algún juego no especificado en función de una serie de atributos que recogen aspectos meteorológicos: pronóstico, temperatura (en grados Fahrenheit), humedad y viento. Esta pequeña base de datos, completamente ficticia, se

utiliza con frecuencia para ilustrar los métodos de aprendizaje automático. Se conoce como «el problema del tiempo». Con este conjunto de ejemplos trataremos de construir un árbol de decisión.

Tabla 1.1. Conjunto de entrenamiento del problema del tiempo

Pronóstico	Temp. (°F)	Humedad (%)	Ventoso	Clase
Soleado	75	70	Verdadero	Jugar
Soleado	80	90	Verdadero	No jugar
Soleado	85	85	Falso	No jugar
Soleado	72	95	Falso	No jugar
Soleado	69	70	Falso	Jugar
Nuboso	72	90	Verdadero	Jugar
Nuboso	83	78	Falso	Jugar
Nuboso	64	65	Verdadero	Jugar
Nuboso	81	75	Falso	Jugar
Lluvioso	71	80	Verdadero	No jugar
Lluvioso	65	70	Verdadero	No jugar
Lluvioso	75	80	Falso	Jugar
Lluvioso	68	80	Falso	Jugar
Lluvioso	70	96	Falso	Jugar

Fuente: Quinlan (1993).

Puesto que no todos los casos pertenecen a la misma clase, el algoritmo «divide y vencerás» partirá el conjunto en subconjuntos. Aún no hemos discutido de qué manera se escoge una pregunta pero, para este ejemplo, supongamos que la pregunta escogida es «pronóstico», que tiene tres respuestas posibles, «soleado», «nuboso» y «lluvioso». El grupo del medio, formado por aquellos casos para los cuales el pronóstico es nuboso, contiene sólo casos de la clase «jugar», pero los subconjuntos primero y tercero todavía tienen casos pertenecientes a las dos clases. Si el primer subconjunto fuese dividido de acuerdo con una pregunta sobre «humedad», con respuestas «humedad ≤ 75» y «humedad > 75», y el tercer subconjunto de acuerdo con la pregunta «ventoso», con salidas «verdadero» y «falso», cada uno de los subconjuntos contendría ahora casos de una única clase. Las divisiones finales de los subconjuntos y el árbol de decisión correspondiente se muestran en las siguientes figuras (figura 1.8 y figura 1.9).

Figura 1.8. Partición final de casos para el problema del tiempo

Pronóstico = soleado:

Humedad ≤ 75:

Pronóstico	Temp. (°F)	Humedad (%)	Ventoso	Decisión
Soleado	75	70	Verdadero	Jugar
Soleado	69	70	Falso	Jugar

Humedad > 75:

Pronóstico	Temp. (°F)	Humedad (%)	Ventoso	Decisión
Soleado	80	90	Verdadero	No jugar
Soleado	85	85	Falso	No jugar
Soleado	72	95	Falso	No jugar

Pronóstico = nuboso:

Pronóstico	Temp. (°F)	Humedad (%)	Ventoso	Decisión
Nuboso	72	90	Verdadero	Jugar
Nuboso	83	78	Falso	Jugar
Nuboso	64	65	Verdadero	Jugar
Nuboso	81	75	Falso	Jugar

Pronóstico = lluvioso:

Ventoso = verdadero:

Pronóstico	Temp. (°F)	Humedad (%)	Ventoso	Decisión
Lluvioso	71	80	Verdadero	No jugar
Lluvioso	65	70	Verdadero	No jugar

Ventoso = falso:

Pronóstico	Temp. (°F)	Humedad (%)	Ventoso	Decisión
Lluvioso	75	80	Falso	Jugar
Lluvioso	68	80	Falso	Jugar
Lluvioso	70	96	Falso	Jugar

Fuente: Quinlan (1993).

Figura 1.9. Árbol de decisión correspondiente al problema del tiempo

```
Pronóstico = soleado:
|   Humedad ≤ 75: Jugar
|   Humedad > 75: No jugar
Pronóstico = nuboso: Jugar
Pronóstico = lluvioso:
|   Ventoso = verdadero: No jugar
|   Ventoso = falso: Jugar
```

Fuente: Quinlan (1993).

1.2.1.1. Criterio de partición

Cualquier pregunta que divida el conjunto de casos de entrenamiento en al menos dos subconjuntos no vacíos resultará finalmente en una partición en subconjuntos de una única clase, incluso aunque todos o la mayoría de ellos sólo contengan un caso. Sin embargo, el objetivo del proceso de construcción de un árbol de decisión es construir un árbol que revele la estructura del dominio, información interesante a la hora de hacer predicciones. Para ello, es necesario un número significativo de casos en cada hoja o, dicho de otro modo, la partición debe tener el menor número de bloques posible. Lo ideal sería escoger una pregunta, en cada nivel, de tal forma que conduzca a que el árbol final sea pequeño.

Ya que se busca la obtención de un árbol de decisión compacto, podríamos pensar en explorar todos los árboles posibles y seleccionar el más simple. Pero, desafortunadamente, el problema de encontrar el árbol de decisión más pequeño consistente con el conjunto de entrenamiento es, como se dice en la jerga computacional, un problema NP completo (no resoluble en tiempo polinomial); habría que examinar un número asombroso de árboles. Por ejemplo, hay más de 4×10^6 árboles de decisión que se derivan del pequeño conjunto de entrenamiento del ejemplo anterior.

Por otro lado, dado que la mayoría de métodos de construcción de árboles de decisión, incluido éste, son, como hemos señalado anteriormente, algoritmos *greedy*, que una vez realizan una selección no vuelven a explorar las consecuencias de selecciones alternativas, es obviamente importante conseguir una buena elección de la partición del conjunto de casos de entrenamiento.

En los experimentos CLS de Hunt se consideraban varias posibilidades para evaluar una pregunta. La mayoría de ellas se basaban en el criterio de la frecuencia de la clase. Por ejemplo, uno de sus programas se restringía a problemas con dos clases, positivo y negativo, siendo preferibles aquellas preguntas en las que el subconjunto de casos asociados a la respuesta contuviera:

- sólo casos positivos; o si no,
- sólo casos negativos; o si no,
- el mayor número de casos positivos.

A pesar de que sus programas usaban criterios simples de este tipo, Hunt sugirió que un enfoque basado en la Teoría de la Información podría proporcionar ventajas (HUNT *et al*., 1966).

El criterio utilizado por Quinlan para hacer las particiones se apoya precisamente en una serie de conceptos procedentes de dicha teoría. Antes de pasar a explicar el criterio empleado por Quinlan, haremos un breve repaso de tales conceptos siguiendo a Reza (1994).

La recepción de un mensaje relativo a un hecho incierto o desconocido proporciona información que podría ser medida como el inverso de la probabilidad del mensaje, ya que un mensaje proporciona más información cuanto más improbable es, es decir, cuando indica que ha ocurrido un suceso poco esperado. No obstante, dado que un mensaje proporciona información nula cuando su probabilidad es 1, esto es, cuando indica que ha ocurrido un suceso cierto, resulta más adecuado usar como medida de la información proporcionada por el mensaje el logaritmo del inverso de su probabilidad:

$$\text{Información del mensaje } i = \log \frac{1}{p_i} = -\log p_i.$$

De este modo, la información será nula cuando la probabilidad del mensaje sea 1.

Es posible utilizar distintas bases para el logaritmo, aunque resulta habitual usar base 2 para, de esta manera, medir la información en bits. Si usásemos el logaritmo natural, la información se mediría en *nats*, y se mediría en *hartleys* si empleásemos logaritmos decimales.

Por ejemplo, si tenemos 8 mensajes equiprobables, la información proporcionada por cada uno de ellos es $-\log_2 \frac{1}{8}$ ó 3 bits.

El promedio de esta magnitud para el conjunto de posibles mensajes $x_1, x_2, \ldots, x_n$ con probabilidades $p_1, p_2, \ldots, p_n$ o, dicho de otro modo, para todas las posibles ocurrencias de una variable aleatoria X, recibe el nombre de *entropía* de la fuente de mensajes, de la variable aleatoria X o de su vector de probabilidades, indistintamente:

$$H(X) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$$

$$X = \{x_1, x_2, \ldots, x_n\}$$

$$p_i = prob\,(X = x_i) \;/\; p_i \geq 0 \text{ y } \sum_{i=1}^{n} p_i = 1.$$

La entropía mide la cantidad de información que obtenemos en promedio cuando recibimos un mensaje o conocemos el valor adoptado por la variable aleatoria.

Así, por ejemplo, la entropía de una moneda será:

$$X = \left\{ \begin{array}{ll} \text{cara} = x_1 & \text{cruz} = x_2 \\ p_1 = \frac{1}{2} & p_2 = \frac{1}{2} \end{array} \right\}$$

$$H(X) = \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1 \text{ bit.}$$

Es decir, observando el resultado (cara o cruz) se recibe 1 bit de información.

La entropía de un dado fiel será:

$$X = \left\{ \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{array} \right\}$$

$$H(X) = \frac{1}{6}\log_2 6 + \ldots + \frac{1}{6}\log_2 6 = \log_2 6 = 2{,}58 \text{ bits.}$$

Y, lógicamente, un resultado cierto no tiene entropía:

$$X = \left\{ \begin{array}{cc} x_1 & x_2 \\ 0 & 1 \end{array} \right\}$$

$$H(X) = 0.$$

La entropía es, por tanto, una medida de la aleatoriedad o incertidumbre de X o de la cantidad de información que, en promedio, nos proporciona el conocimiento de X.

La entropía de una variable aleatoria será mayor cuanta más incertidumbre exista con respecto al valor que adoptará.

Consideremos una variable aleatoria X cuyo rango de valores es $x_1, x_2,..., x_n$ y su distribución de probabilidades es $p_1, p_2,..., p_n$. Se cumple que:

$$H(1, 0,...,0) \leq H(X) \leq H\left(\frac{1}{n},..., \frac{1}{n}\right).$$

Es decir, la entropía de X está acotada entre la entropía del vector de probabilidades $(1, 0,..., 0)$ y la del vector de probabilidades $\left(\frac{1}{n},..., \frac{1}{n}\right)$ o, dicho de otro modo, la entropía de una variable aleatoria será mínima cuando su valor sea cierto y máxima cuando todos los valores sean equiprobables.

Como $H(1, 0,...,0) = 0$ y $H\left(\frac{1}{n},..., \frac{1}{n}\right) = \sum_{i=1}^{n} \frac{1}{n} \log_2 n = \log_2 n,$

entonces $\boxed{0 \leq H(X) \leq \log_2 n}$.

Como se puede observar, la entropía de una fuente de mensajes equiprobables crece con n, luego cuantos más valores equiprobables pueda adoptar X mayor será la entropía de dicha variable aleatoria o de la fuente de mensajes. Así, por ejemplo, la entropía de un dado fiel (2,58 bits) es mayor que la de una moneda (1 bit).

Análogamente, podemos definir la «entropía conjunta» de dos variables aleatorias X e Y. Sea un suceso que viene definido por las variables X e Y con rangos de valores $\{x_1,..., x_r\}$, $\{y_1,..., y_s\}$ y distribución de probabilidad

$$p(x_i, y_j) = p\{X = x_i; Y = y_j\}$$

$$i = 1,..., r; j = 1,..., s; \sum_i \sum_j p(x_i, y_j) = 1; p(x_i, x_j) \geq 0,$$

se define la entropía conjunta de X e Y como:

$$H(X,Y) = \sum_i \sum_j p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)}.$$

En lo sucesivo, para simplificar la notación, eliminaremos los subíndices, escribiendo ahora la entropía conjunta de dos variables aleatorias X e Y como:

$H(X,Y) = \sum_x \sum_y p(x, y) \log_2 \frac{1}{p(x, y)}$, siendo $p(x,y)$ la probabilidad conjunta: $p(x,y) = p\{X = x; Y = y\}$.

La entropía conjunta será la cantidad de información que obtendremos, en promedio, al conocer los valores x e y que adoptan las variables aleatorias X e Y. Por ejemplo, en un problema de clasificación, la variable aleatoria X podría representar la clase de un elemento y la variable Y uno de los atributos.

Siempre se cumple que $H(X,Y) \leq H(X) + H(Y)$, verificándose la igualdad si y sólo si X e Y son variables aleatorias independientes, es decir, si $p(x,y) = p(x).p(y) \cup x, y$. Por tanto, la entropía conjunta de dos variables aleatorias independientes será la suma de entropías, y si las variables aleatorias son dependientes, será menor que la suma de entropías. Generalizando a más variables:

$$H(X_1, X_2, X_3, \ldots) \leq H(X_1) + H(X_2) + H(X_3) + \ldots$$

Definamos ahora la «entropía condicional» de X dada Y. $H(X)$ es la incertidumbre que se tiene sobre X antes de conocer Y. Supongamos que sucede Y y que toma el valor y. Entonces los valores $p(x|y)$ representan la probabilidad de X dado que $Y = y$, y por lo tanto la entropía remanente de X es: $H(X|y) = \sum_x p(x|y) \log_2 \frac{1}{p(x|y)}$. Esta cantidad $H(X|y)$ es a su vez una variable aleatoria definida en el rango de Y, entonces la «entropía condicional» $H(X|Y)$ será la esperanza de esta variable aleatoria:

$$H(X|Y) = E_Y(H(X|Y = y)) = \sum_y p(y) \cdot H(X|Y = y) =$$

$$\sum_y p(y) \sum_x p(x|y) \log_2 \frac{1}{p(x|y)}.$$

Como la probabilidad condicional es: $p(x|y) = p\{X=x|Y=y\} = \frac{p(x,y)}{p(y)}$, la igualdad anterior queda:

$$H(X|Y) = \sum_y p(y) \sum_x \frac{p(x,y)}{p(y)} \log_2 \frac{1}{p(x,y)} = \boxed{\sum_x \sum_y p(x,y) \log_2 \frac{1}{p(x,y)}}.$$

La entropía condicional de X dada Y, $H(X|Y)$, es una medida de la incertidumbre respecto a X cuando se conoce el valor de Y, es decir, es la cantidad de información que necesitamos para conocer plenamente X cuando ya tenemos la información suministrada por Y. Se cumple que:

- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.
- $H(X|Y) \leq H(X)$, verificándose la igualdad si y sólo si X e Y son variables aleatorias independientes.

Tenemos entonces que $H(X)$ es la incertidumbre sobre X y que $H(X|Y)$ es la incertidumbre sobre X cuando conocemos el valor de Y, que será menor, pues al conocer Y tenemos más información que nos puede ayudar a reducir la incertidumbre sobre X. A esa reducción de incertidumbre se la denomina «información mutua» entre X e Y:

$$I(X;Y) = H(X) - H(x|Y),$$

que es la información que una de las variables nos transmite sobre la otra (que sólo será nula cuando las variables sean independientes).

Se cumple además que $I(X;Y) = I(Y;X)$, puesto que, como ya hemos indicado, $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$, y, por tanto, $H(X) - H(X|Y) = H(Y) - H(Y|X)$, es decir, $I(X;Y) = I(Y;X)$.

Generalizando a más variables, $I(X;Y,Z) = H(X) - H(X|Y,Z)$, y se cumple que $I(X;Y) \leq I(X;Y,Z)$, ya que Y y Z proporcionarán más información acerca de X de la que lo harían sólo Y o sólo Z, verificándose la igualdad cuando $p(x|y,z) = p(x|y)$, es decir, cuando dado Y, sean independientes X y Z.

Para nuestro problema de clasificación, podemos considerar que X es una variable aleatoria que nos muestra la clase a la que pertenece un elemento, mientras que Y_i, $i = 1, 2, \ldots, n$, son los atributos o variables que caracterizan a los elementos que queremos clasificar. Así, en nues-

tro estudio empírico, X indicará si la empresa pertenece al grupo de las sanas o al grupo de las fracasadas, mientras que las variables Y_i serán los ratios financieros utilizados para la clasificación.

El criterio empleado en el algoritmo ID3 —predecesor del C4.5— para seleccionar el atributo en base al cual realizar una partición se basa en la información mutua entre el atributo y la clase. Quinlan lo denomina «criterio de ganancia» *(gain criterion)*. Consiste en seleccionar para hacer cada partición aquel atributo Y_i que proporcione la máxima información sobre X, es decir, que maximiza $I(X;Y_i)$, magnitud a la que Quinlan denomina «ganancia de información».

Para ilustrar esto, consideremos de nuevo el problema del tiempo de la página 30. Hay dos clases; 9 casos pertenecen a la clase «jugar» y 5, a la clase «no jugar». Entonces,

$$H(X) = \frac{9}{14} \cdot \log_2 \frac{1}{\frac{9}{14}} + \frac{5}{14} \cdot \log_2 \frac{1}{\frac{5}{14}} = 0{,}940 \text{ bits.}$$

El resultado de utilizar el atributo «pronóstico» para dividir el conjunto de entrenamiento T en tres subconjuntos es:

$$H(X|Y_1) = E_{Y1}(H(X|Y_1 = y_1)) = 5/14 \cdot (-2/5 \cdot \log_2(2/5) - 3/5 \cdot \log_2(3/5)) + 4/14 \cdot (-4/4 \cdot \log_2(4/4) - 0/4 \cdot \log_2(0/4)) + 5/14 \cdot (-3/5 \cdot \log_2(3/5) - 2/5 \cdot \log_2(2/5)) = 0{,}694 \text{ bits.}$$

La «ganancia de información» proporcionada por el atributo «pronóstico» (o lo que es lo mismo, la información mutua entre el atributo y la clase) es, por tanto, $H(X) - H(X|Y_1) = 0{,}940 - 0{,}694 = 0{,}246$ bits.

Supongamos ahora que en lugar de dividir el conjunto T de acuerdo con el atributo «pronóstico», lo partimos de acuerdo con el atributo «ventoso». Esto dará lugar a dos subconjuntos, uno con 3 casos de la clase «jugar» y 3 de la clase «no jugar», y el otro con 6 casos de la clase «jugar» y 2 de la clase «no jugar». Por tanto,

$$H(X|Y_2) = 6/14 \cdot (-3/6 \cdot \log_2(3/6) - 3/6 \cdot \log_2(3/6)) + 8/14 \cdot (-6/8 \cdot \log_2(6/8) - 2/8 \cdot \log_2(2/8)) = 0{,}892 \text{ bits.}$$

Y la ganancia sería entonces de 0,940 – 0,892 = 0,048 bits, menor que la ganancia resultante anterior. Así, el criterio de

ganancia preferiría la pregunta sobre «pronóstico» frente a la de «ventoso».

A pesar de que este criterio proporciona buenos resultados, introduce un sesgo en favor de las preguntas con muchas respuestas posibles, esto es, los atributos con muchos valores distintos. Como indica Quinlan (1993), esto puede verse considerando un problema hipotético de diagnóstico médico en el que uno de los atributos contiene la identificación del paciente. Como tal identificación será distinta para cada paciente, la partición de cualquier conjunto de casos de entrenamiento de acuerdo con este atributo conducirá a un gran número de subconjuntos, y cada uno de ellos contendrá sólo un caso. Al sólo tener un caso, cada uno de estos subconjuntos necesariamente contendrá casos de una única clase y, por tanto, $H(X|Y) = 0$, con lo que la ganancia de información obtenida al usar este atributo para partir el conjunto de casos de entrenamiento sería máxima. Pero, sin embargo, desde el punto de vista de la predicción, tal partición sería completamente inútil.

Para corregir este sesgo, el algoritmo C4.5 utiliza como criterio de partición el consistente en seleccionar aquel atributo Y_i que maximice la magnitud $\frac{I(X;Y_i)}{H(Y_i)}$, a la que Huinlan denomina «ratio de ganancia» *(gain ratio)*. Cuando tenga muchos valores posibles, $I(X;Y_i)$, será grande, pero también lo será, en general, su entropía $H(Y_i)$, y, por tanto, se compensará un elemento con otro.

Como $I(X;Y_i)$ es la información que Y_i nos proporciona sobre X, y $H(Y_i)$ es la información que obtenemos al conocer la variable Y_i, resulta entonces que el ratio de ganancia es el porcentaje de la información proporcionada por Y_i que es útil para conocer X.

Adicionalmente, para evitar la selección de un atributo simplemente porque su entropía $H(Y_i)$ sea pequeña, lo que aumentaría el valor del cociente anterior, se exige además que $I(X;Y_i)$ sea razonablemente grande, al menos del orden de la información mutua media para todos los atributos examinados.

De este modo, volviendo al ejemplo del diagnóstico médico, dividir el conjunto de casos de entrenamiento de acuerdo con el criterio de ganancia en base al atributo que contiene la identificación del paciente,

conduciría a la máxima ganancia por ser la entropía condicional nula, y esta ganancia, que sería la entropía del conjunto de casos de entrenamiento, tendrá como máximo un valor de $\log_2 k$, siendo k el número de clases. Pero al añadir ahora el denominador $H(Y_i)$, que sería $\log_2 n$, siendo n el número de casos de entrenamiento (ya que el atributo tendrá tantas respuestas equiprobables como pacientes existan), parece razonable presumir que el número de casos de entrenamiento será mucho mayor que el número de clases, así que el ratio adoptará un pequeño valor y, en consecuencia, este atributo no sería seleccionado para llevar a cabo la partición.

Continuando con el ejemplo de la página 30 (el problema del tiempo), la pregunta sobre «pronóstico» da lugar a 3 subconjuntos que contienen 5, 4 y 5 casos, respectivamente. $H(Y_i)$ sería:

$$H(Y_1) = 5/14 \cdot \log_2(5/14) - 4/14 \cdot \log_2(4/14) - 5/14 \cdot \log_2(5/14) = 1{,}577 \text{ bits.}$$

Para este atributo, cuya información mutua o ganancia es, como vimos previamente, 0,246 bits, el ratio de ganancia es 0,246 / 1,577 = 0,156.

En palabras de Quinlan (1993), «el criterio del ratio de ganancia es robusto y generalmente proporciona una mejor selección de la pregunta que el criterio de ganancia».

1.2.1.2. Particiones posibles consideradas

En resumen, el algoritmo C4.5 emplea un enfoque *divide-and-conquer* para generar un árbol de decisión a partir de un conjunto D de casos del modo siguiente:

- Si D satisface un «criterio de parada», el árbol para dicho conjunto será una hoja asociada con la clase más frecuente en el mismo. Una razón para detener la ramificación del árbol es que D contenga sólo casos de la misma clase.
- Alguna pregunta T acerca de un atributo con respuestas mutuamente excluyentes $T_1, T_2, \ldots, T_k$, seleccionada de acuerdo con el criterio de partición visto anteriormente, se utiliza para dividir D

en los subconjuntos $D_1, D_2, \ldots, D_k$, donde D_i contiene aquellos casos con respuesta T_i. El árbol para D contiene la pregunta T como nodo raíz con un subárbol para cada respuesta T_i que se construye aplicando recursivamente el mismo procedimiento a los casos de D_i.

En algunas situaciones, cada pregunta posible divide D en subconjuntos con la misma distribución de clases y, por tanto, todas ellas proporcionan ganancia cero; C4.5 tiene en cuenta esto como un criterio de parada adicional.

Dado que los árboles más pequeños son preferibles, por ser más fáciles de entender y a menudo predictores más precisos, se examina una familia de preguntas posibles y se escoge una de ellas para maximizar el valor del criterio de partición.

Los atributos que describen los casos pueden ser clasificados en atributos continuos, cuyos valores son numéricos, y atributos discretos, con valores nominales no ordenados. Por ejemplo, la descripción de una persona podría incluir su peso en kilogramos, con un valor tal como 71,2, y el color de sus ojos, cuyo valor podría ser «azul», «marrón», etc.

La mayoría de sistemas de construcción de clasificadores definen un tipo de preguntas posibles y examinan todas las preguntas de ese tipo. Convencionalmente, una pregunta involucra a sólo un atributo, porque de este modo el árbol es más fácil de entender y elude la explosión combinatoria que resultaría si atributos múltiples pudiesen aparecer en una misma pregunta. Los tipos de pregunta considerados por C4.5 son:

- $A = ?$ para un atributo discreto A, con una respuesta y rama para cada posible valor de dicho atributo.
- Aunque la anterior es la opción por defecto implementada en C4.5 para los atributos discretos, también se contempla la posibilidad de agrupar los valores de un atributo discreto en un número variable de grupos con una respuesta para cada grupo en vez de para cada valor del atributo. Si bien el enfoque estándar basado en generar una rama separada y un subárbol para cada uno de los posibles valores del atributo es generalmente apropiado (puesto que las diferencias en la distribución de clases que conducen a la selección de este atributo sugieren que los casos de entrenamiento

asociados con cada valor del atributo debieran ser tratados separadamente), si existen muchos valores, sin embargo, este enfoque conlleva dos problemas. Por un lado, una consecuencia de dividir el conjunto de entrenamiento en numerosos subconjuntos es que cada uno de ellos sea pequeño, y los patrones útiles en los subconjuntos podrían ser indetectables debido a la insuficiencia de datos. Por otro lado, aunque la evidencia empírica sugiere que generalmente el criterio del ratio de ganancia proporciona buenas selecciones del atributo de partición, debido a que el denominador de dicho ratio de ganancia crece rápidamente a medida que aumenta el número de subconjuntos, puede que en ocasiones resulten excesivamente penalizados los atributos con muchos valores, lo que indicaría la conveniencia de agrupar estos valores.

C4.5 utiliza otro algoritmo *greedy* para llevar a cabo dicha agrupación. Esto tiene lugar a través de un proceso iterativo de fusión de grupos de valores, tomando como grupos iniciales justamente los valores individuales del atributo considerado. En cada iteración, dos de los grupos de valores existentes en ese momento son fusionados (reduciéndose en una unidad el número de grupos). Para decidir cuál será la pareja de grupos que se fusionará se ensaya con todas las posibles parejas (manteniendo en cada ensayo inalterados los restantes grupos) y se selecciona aquella pareja cuya fusión conduce al máximo valor del criterio de partición, *gain* o *gain ratio* (siempre que el valor de esta magnitud aumente respecto a la situación de partida). Sucesivas iteraciones van reduciendo el número de grupos hasta que sólo queden dos o no sea posible mejorar el criterio de partición.

Como acabamos de señalar, la partición que surge de cualquier colección particular de grupos de valores es evaluada a través del criterio de partición. Si se utiliza el ratio de ganancia, el criterio implementado por defecto en C4.5, la fusión de grupos de valores de atributos resultará en menos subconjuntos de casos de entrenamiento y en la correspondiente reducción de la entropía del atributo (el denominador del ratio de ganancia). Si la reducción en la ganancia de información (el numerador del ratio de ganancia) no es

sustancial, el ratio de ganancia final podría aumentar. Sin embargo, si se utiliza el criterio de ganancia (opcional), la fusión repetida siempre producirá una partición con menor ganancia (ya que la entropía del conjunto condicionada al atributo será mayor al tener éste menos salidas y, por tanto, la ganancia o información mutua será menor). La fusión en este caso por sí sola no se producirá, por lo que deberá ser forzada hasta que sólo queden dos grupos de valores.

– Para los atributos A continuos, la pregunta será binaria con respuestas $A \leq Z$ y $A > Z$, basada en comparar el valor de A con un valor umbral Z que maximice el criterio de partición. Para encontrar dicho umbral Z, los casos del conjunto de entrenamiento se clasifican según los valores que toman para el atributo A considerado, se ordenan dichos valores, eliminando los repetidos, y se calcula el valor intermedio entre cada par de valores. Por ejemplo, si el conjunto de entrenamiento consta de 10 casos que toman los valores para el atributo A {0.2, 0.3, 0.7, 0.1, 0.8, 0.45, 0.33, 0.1, 0.8, 0}, el conjunto de valores ordenados sería {0, 0.1, 0.2, 0.3, 0.33, 0.45, 0.7, 0.8} y el conjunto finito de posibles umbrales sería {0.05, 0.15, 0.25, 0.315, 0.39, 0.575, 0.75}. Con estos siete valores, tendremos siete particiones posibles $(A \leq 0.05, A > 0.05)$, $(A \leq 0.15, A > 0.15)$, $(A \leq 0.25, A > 0.25)$, etc.

Denotando los valores distintos ordenados del atributo A por $\{v_1, v_2, \ldots, v_m\}$ cualquier valor umbral entre v_i y v_{i+1} causará el mismo efecto que dividir los casos en aquéllos que toman valores para el atributo A comprendidos en $\{v_1, v_2, \ldots, v_i\}$ y aquéllos que toman valores comprendidos en $\{v_{i+1}, v_{i+2}, \ldots v_m\}$. Cada par de valores adyacentes sugiere un umbral potencial $Z = (v_i + v_{i+1})/2$, y una correspondiente partición del conjunto de entrenamiento. Existirán, por tanto, $m - 1$ particiones posibles para el atributo A que habrán de ser examinadas. El umbral que proporcione el mejor valor del criterio de partición será entonces seleccionado para el atributo A.

Si bien sería computacionalmente costoso examinar todos los $m - 1$ posibles umbrales, Fayyad e Irani (1992) demuestran que, para criterios de partición convexos, tales como la ganancia de información, no es necesario examinar todos esos umbrales, ya que si todos los casos con valor v_i o

v_{i+1} pertenecen a la misma clase, un umbral entre ellos no puede conducir a una partición que proporcione el máximo valor para el criterio.

Aunque la mayoría de algoritmos seleccionan el punto medio de un intervalo como el umbral representativo, es decir, $\frac{v_i + v_{i+1}}{2}$, C4.5 selecciona como umbral el mayor valor de A en el conjunto de entrenamiento que no exceda dicho punto medio en lugar del punto medio en sí, con lo que se asegura que los valores umbrales que aparecen en el árbol de decisión realmente se dan en los datos.

En definitiva, cuando los atributos sean continuos el valor del criterio de partición será función del umbral. La posibilidad de seleccionar el umbral Z que maximice dicho valor proporciona a un atributo continuo A una ventaja sobre un atributo discreto, donde no existen tales umbrales, y también sobre otros atributos continuos que tomen menos valores distintos en el conjunto de entrenamiento. Pensemos que, cuantos más umbrales se puedan evaluar, más posibilidades habrá de que alguno de ellos maximice el valor del criterio de partición, y, por tanto, los atributos continuos con numerosos valores distintos resultarán favorecidos en el proceso de selección de la pregunta a considerar en un nodo.

La última versión del algoritmo C4.5 (C4.5 *Release* 8) incluye una modificación del mismo realizada para corregir el sesgo al que acabamos de referirnos (Quinlan, 1996). Dicha modificación, inspirada en el principio de la «longitud de descripción mínima» o principio MDL —*Minimum Description Length*— (Rissanen, 1983), ajusta la ganancia de información proporcionada por una pregunta sobre un atributo continuo.

En primer lugar, introduzcamos brevemente el principio MDL. Siguiendo a Quinlan y Rivest (1989), supongamos un emisor y un receptor que poseen una lista ordenada de valores de atributos de un conjunto de casos de entrenamiento. El emisor, además, conoce la clase a la que pertenece cada caso y debe transmitir esta información al receptor. El emisor debe entonces codificar y enviar una teoría acerca de cómo clasificar los casos y, puesto que la teoría podría ser imperfecta, también debe identificar las excepciones a la misma que se dan en los casos de entrenamiento y establecer cómo deberían corregirse las clases de los mismos predichas por la teoría. La longitud total de la transmisión será por tanto el número de bits requeridos para codificar la teoría (el «coste de la

teoría») más el número de bits necesarios para identificar y corregir las excepciones (el «coste de las excepciones»). El emisor deberá seleccionar una teoría entre varias alternativas, siendo algunas simples pero con muchas excepciones y otras más elaboradas y complejas pero más precisas, es decir, con menos excepciones. Según establece el principio MDL, habría que seleccionar aquella teoría que minimice la suma de los costes de la teoría y de las excepciones. Más adelante formalizaremos este principio.

El coste de las excepciones asociado con un conjunto de casos D es equivalente a $|D| \times H(D)$, siendo $|D|$ el número de casos del conjunto D y $H(D)$, la entropía de dicho conjunto, en nuestro caso, la cantidad de información en promedio necesaria para identificar la clase a la que pertenece un elemento del conjunto. Entonces $|D| \times Gain(D,T)$ (recordemos que $Gain(D,T)$ será la información mutua entre D y T, es decir, la reducción de la incertidumbre acerca de la clase o la ganancia de información que se consigue al dividir el conjunto D de acuerdo con la pregunta T mide la reducción en el coste de las excepciones cuando D es dividido de acuerdo con la pregunta T. La partición del conjunto D de esta manera, sin embargo, requiere la transmisión de una teoría más compleja que incluya la definición de T. Mientras que una pregunta $A = ?$ sobre un atributo discreto A puede especificarse simplemente nombrando al atributo en cuestión (por ejemplo, «color de ojos»), una pregunta $A \leq Z$ también incluye al umbral Z; si hay N valores distintos del atributo A en el conjunto de casos y, por tanto, $N-1$ posibles umbrales para A, se requerirán $\log_2 (N-1)$ bits adicionales. Como señala Quinlan (1996), «incluso con un criterio de partición convexo que satisface los requisitos de Fayyad e Irani (1992), no podemos usar el número N' de umbrales potencialmente maximizadores de la ganancia en vez del número mayor N de umbrales posibles. Dado que el receptor conoce los valores de los atributos de los casos pero no sus clases, no puede determinar si todos los casos con dos valores adyacentes de A pertenecen a la misma clase. El mensaje debe identificar consecuentemente el umbral escogido entre todos los umbrales posibles».

La modificación consiste en «cobrarle» este coste creciente asociado con una pregunta sobre un atributo continuo a la ganancia alcanzada por la misma, es decir, reducir la ganancia de información (por caso)

$Gain(D,T)$ en $\log_2 (N - 1)/|D|$. De este modo será menos probable que una pregunta sobre un atributo continuo con numerosos valores distintos proporcione el máximo valor para el criterio de partición de entre la familia de preguntas posibles y, por tanto, será menos probable que sea seleccionada. Además, si todos los umbrales sobre un atributo continuo A proporcionaran una ganancia ajustada menor que cero, el atributo A sería efectivamente descartado. Las consecuencias de esta modificación son entonces un nuevo *ranking* de preguntas potenciales y la posible exclusión de alguna de ellas.

1.2.1.3. Valores faltantes

Un aspecto importante a tener en cuenta cuando se lleva a cabo la aplicación de alguna técnica de minería de datos es qué hacer cuando se desconoce el valor que adopta alguno de los atributos para un caso concreto, es decir, cuando aparecen valores perdidos *(missing values)*. En el pequeño conjunto de entrenamiento del problema del tiempo de la página 30, el valor que toma cada uno de los cuatro atributos predictores es conocido para los 14 casos, pero esto no siempre tiene por qué ser así.

La mayoría de técnicas de minería de datos requieren para su aplicación el conocimiento de la totalidad de los valores de los atributos para todos los ejemplos, esto es, no aceptan la existencia de vectores incompletos en la base de datos usada para el entrenamiento, para estimar el modelo. La alternativa más simple para convertir una base de datos incompleta en una completa consistiría en eliminar aquellos casos de la base de datos para los que el valor de algún atributo es desconocido. Pero no son poco habituales las situaciones en las que el conjunto de entrenamiento es pequeño, y ello implica el aprovechamiento de ejemplos que podrían ser desechados si dispusiésemos de una gran cantidad de casos. Pensemos en que estos casos para los que el valor de algún atributo es desconocido pueden esconder en el resto de sus atributos información relevante de cara a la detección de patrones útiles a partir de los datos, y resultaría entonces más adecuado completar esos vectores de la base de datos en vez de eliminarlos. Existen varias for-

mas de completar dichos casos. La primera de ellas consiste en asignar a un atributo ausente de un caso el resultado de la media de los valores que toma ese atributo en los casos de entrenamiento en los que sí está presente, aunque obviamente se esté sesgando el caso. Otra forma de completar los casos sería empleando, en lugar de la media, la mediana, o también se podría utilizar algún otro método más sofisticado de imputación de valores desconocidos.

C4.5 implementa su propio procedimiento para poder trabajar con conjuntos de entrenamiento que contengan *missing values*. En primer lugar, recordemos que C4.5 utiliza como criterio de partición el consistente en seleccionar aquel atributo Y_i que maximice el ratio de ganancia, $\frac{I(X;Y_i)}{H(Y_i)}$, luego, lógicamente, para poder seleccionar un atributo habrá que calcular esta magnitud para todos ellos. Supongamos que queremos calcular el ratio de ganancia para un atributo Y_i, así que habrá que hallar:

$$I(X;Y_1) = H(X) - H(X|Y_1).$$

Dado que $H(X|Y_1) = E_{Y_1}(H(X|Y_1 = y_1)) = \sum_{y_1} p(y_1) \cdot H(X|Y_1 = y_1)$, también podemos expresar la información mutua como:

$$I(X;Y_1) = \sum_{y_1} p(y_1)\,[H(X) - H(X|Y_1 = y_1)].$$

Si algún valor del atributo Y_i es desconocido, se asumirá que este hecho no proporciona ninguna información sobre X, de modo que admitimos que $H(X|Y_1 = desconocido) = H(X)$, y el sumando correspondiente de la expresión anterior será nulo. Por tanto,

$$I(X;Y_1) = \sum_{y_1 \neq desconocido} p(y_1)\,[H(X) - H(X|Y_1 = y_1)].$$

En la implementación de Quinlan, el cálculo de los valores $p(y_1)$ se realizará del modo usual, es decir, número de casos en los que $Y_1 = y_1$ entre número de casos totales (incluyendo los casos con valor desconocido para ese atributo Y_i). Sin embargo, $H(X)$ debe ser calculada teniendo en cuenta exclusivamente aquellos casos para los cuales el valor del atributo Y_1 es conocido.

En cuanto al denominador del ratio de ganancia, esto es, $H(Y_i)$, también se calculará de la manera habitual, considerando $y_1 = desconocido$ como un valor posible más que puede tomar Y_1:

$$H(Y_1) = \sum_{y_1} p(y_1) \cdot \log_2 \frac{1}{p(y_1)}.$$

De este modo, queda resuelto el problema del cálculo del ratio de ganancia necesario para llevar a cabo la selección de los atributos. Pero una vez seleccionado el atributo en base al cual realizar una partición, se nos plantea un nuevo problema: cómo realizar la partición de los casos de entrenamiento que toman un valor desconocido para ese atributo seleccionado. Supongamos que llegamos a un nodo que contiene una pregunta acerca de cierto atributo, y alguno de los casos toma un valor desconocido para el mismo, es decir, no tiene respuesta a dicha pregunta, así que no podría ser asignado a un subconjunto particular. Entonces este caso con valor desconocido para ese atributo se distribuye probabilísticamente de acuerdo con la frecuencia relativa de los casos con respuestas conocidas. Si, por ejemplo, hay 10 casos con respuesta conocida, de los cuales 6 se asignan a una rama y 4, a otra, los casos con respuesta desconocida se asignarían a esas dos ramas en proporciones 6/10 y 4/10, respectivamente. De este modo, en los nodos hoja del árbol aparecerán valores fraccionarios. Así, al concluir el proceso, a cada caso con algún atributo desconocido le corresponderá una distribución de probabilidad sobre las distintas clases. Finalmente, la clase que se asigna al caso será la más probable de acuerdo con dicha distribución.

Para clarificar esto, volvamos al problema del tiempo, e imaginémonos ahora que el valor del atributo «pronóstico» para el sexto caso de la tabla 1.1 es desconocido. El ratio de ganancia para este atributo ya no sería el mismo de antes, sino que disminuiría. Aun así, supongamos que resulta seleccionado. Cuando los 14 casos de entrenamiento se dividen de acuerdo con esta pregunta, los 13 casos para los cuales el valor de «pronóstico» es conocido no suponen ningún problema. El otro caso se asignaría a todos los bloques de la partición, correspondientes a las respuestas «soleado», «nuboso» y «lluvioso»,

en proporciones o pesos 5/13, 3/13 y 5/13, respectivamente. Así, si tomamos por ejemplo el subconjunto correspondiente a la respuesta «lluvioso», tendremos:

Pronóstico	Temp. (°F)	Humedad (%)	Ventoso	Decisión	Peso
Lluvioso	71	80	Verdadero	No jugar	1
Lluvioso	65	70	Verdadero	No jugar	1
Lluvioso	75	80	Falso	Jugar	1
Lluvioso	68	80	Falso	Jugar	1
Lluvioso	70	96	Falso	Jugar	1
?	72	90	Verdadero	Jugar	5/13

Si este subconjunto fuese dividido de acuerdo con la misma pregunta sobre «ventoso» que en la figura 1.8, las distribuciones de las clases en los subconjuntos serían:

Ventoso = Verdadero:	2 clase No jugar,	5/13 clase Jugar
Ventoso = Falso:	0 clase No jugar,	3 clase Jugar

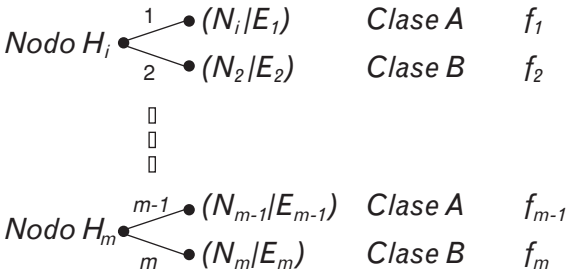
El segundo subconjunto contiene casos de una única clase «jugar», pero el primero todavía contiene casos de ambas clases. No obstante, no es posible encontrar una pregunta que mejore sensiblemente la situación. De manera similar, el subconjunto correspondiente a la respuesta «soleado» tampoco podría ser dividido en subconjuntos de una única clase. Así que el árbol de decisión, como se puede ver en la siguiente figura (figura 1.10), tiene la misma estructura de antes, aunque se han añadido al final de cada hoja una serie de valores de la forma (N) o (N/E). N es la suma de casos que llegan hasta esa hoja y E, el número de ellos que pertenecen a otra clase distinta a la etiquetada en la hoja, es decir, son mal clasificados. Esos números pueden no ser enteros, debido a que los casos que posean valor desconocido para el atributo considerado en un nodo se distribuyen en los distintos subconjuntos de manera fraccionaria. Así, «No jugar (2.4/0.4)» significa que 2.4 casos de entrenamiento (fraccionarios) alcanzan esa hoja, de los cuales 0.4 no pertenecen a la clase «No jugar».

Figura 1.10. Árbol de decisión correspondiente al problema del tiempo considerando valores desconocidos

```
Pronóstico = Soleado:
|   Humedad ≤ 75: Jugar (2.0)
|   Humedad > 75: No jugar (3.4/0.4)
Pronóstico = Nuboso: Jugar (3.2)
Pronóstico = Lluvioso:
|   Ventoso = Verdadero: No jugar (2.4/0.4)
|   Ventoso = Falso: Jugar (3.0)
```

Fuente: Quinlan (1993).

Por último, una vez obtenido el árbol, cuando éste sea usado para clasificar nuevos casos, se plantea el problema de cómo llevar a cabo esta clasificación si el caso tiene valores desconocidos para los atributos considerados en algunos nodos de decisión. La respuesta es sencilla: el caso se va fraccionando de acuerdo con la proporción en la que se han asignado los casos de entrenamiento que llegan a ese nodo entre las distintas ramas que cuelgan de él. Las distintas fracciones del caso continúan avanzando por el árbol (con fraccionamientos sucesivos si atraviesan nuevos nodos con atributos de valor desconocido) hasta llegar a las hojas del árbol. De esta manera el caso quedará en principio repartido entre las distintas hojas en fracciones f_i (cumpliéndose por supuesto que $\sum f_i = 1$). Llegaremos entonces a una situación como la siguiente (en donde se representan los nodos finales y hojas del árbol y, para cada hoja, la clase que se le asigna y la fracción del caso que llega a ella, siendo los valores N_i y E_i, como hemos visto antes, el número de casos de entrenamiento cubiertos por las hojas y el número de ellos mal clasificados):



Una fracción f_1 del caso inicial llega a la hoja *1* (etiquetada como clase A), otra fracción f_2 llega a la hoja *2* (etiquetada como clase B), etc. La hoja *1* clasifica correctamente en la clase A un porcentaje $\frac{N_1 - E_1}{N_1}$ de los casos que llegan a ella, mientras que el porcentaje $\frac{E_1}{N_1}$ restante pertenecen a la clase B. La hoja *2* clasifica correctamente en la clase B el porcentaje $\frac{N_2 - E_2}{N_2}$ de los casos que llegan a ella, mientras que el porcentaje $\frac{E_2}{N_2}$ restante pertenece a la clase A, etc.

Sumando para las distintas hojas obtenemos la parte del caso que debe ser considerada como perteneciente a la clase A (será la probabilidad de que el caso pertenezca a esa clase):

$$p_A = f_1 \cdot \left(\frac{N_1 - E_1}{N_1}\right) + f_2 \cdot \left(\frac{E_2}{N_2}\right) + \ldots + f_{m-1} \cdot \left(\frac{N_{m-1} - E_{m-1}}{N_{m-1}}\right) + f_m \cdot \left(\frac{E_m}{N_m}\right),$$

mientras que la parte que debe ser considerada como perteneciente a la clase B (será la probabilidad de que el caso pertenezca a tal clase) es:

$$p_B = f_1 \cdot \left(\frac{E_1}{N_1}\right) + f_2 \cdot \left(\frac{N_2 - E_2}{N_2}\right) + \ldots + f_{m-1} \cdot \left(\frac{E_{m-1}}{N_{m-1}}\right) + f_m \cdot \left(\frac{N_m - E_m}{N_m}\right).$$

Como resultado de lo anterior, a cada caso con algún atributo desconocido le corresponderá una distribución de probabilidad sobre las distintas clases y el caso será finalmente asignado a la clase más probable de acuerdo con dicha distribución.

Si, por ejemplo, usásemos el árbol de decisión de la figura 1.10 para clasificar un nuevo caso con atributos: «pronóstico = soleado», «temperatura = 60 ºF», «humedad = desconocido» y «ventoso = verdadero», la respuesta a la pregunta «pronóstico» (soleado) indica que debemos asignar el caso al primer subárbol. Pero una vez ahí, la siguiente pregunta es sobre «humedad», para la que este caso no tiene respuesta. Entonces, el caso sería distribuido de la siguiente manera:

• *Jugar:*

$$p_A = \left(\frac{N_1}{N_1+N_2}\right)\cdot\left(\frac{N_1-E_1}{N_1}\right)+\left(\frac{N_2}{N_1+N_2}\right)\left(\frac{E_2}{N_2}\right)=\frac{N_1-E_1+E_2}{N_1+N_2}=\frac{2-0+0\cdot 4}{2+3\cdot 4}=44\,\%$$

• *No jugar:*

$$p_B = \left(\frac{N_2}{N_1+N_2}\right)\cdot\left(\frac{N_2-E_2}{N_2}\right)+\left(\frac{N_1}{N_1+N_2}\right)\left(\frac{E_1}{N_1}\right)=\frac{N_2-E_2+E_1}{N_1+N_2}=\frac{3\cdot 4-0\cdot 4+0}{2+3\cdot 4}=66\,\%$$

y sería asignado a la clase más probable, es decir, a la clase «No jugar».

1.2.1.4. Poda del árbol de decisión

El proceso de construcción del árbol de decisión mediante la aplicación reiterada del procedimiento descrito anteriormente finaliza cuando se alcanza la pureza del nodo, es decir, cuando al nodo sólo le corresponden casos pertenecientes a una de las clases del problema, o cuando la posterior ramificación del árbol ya no suponga ninguna mejora.

Generalmente, este método recursivo de construcción de árboles de decisión conducirá a la generación de árboles muy complejos y excesivamente ajustados a los datos del conjunto utilizado para dicha construcción. En consecuencia, harán una clasificación cuasi-perfecta. Esto, que en principio puede parecer óptimo, en realidad no lo es, ya que ajustarse demasiado a los datos de entrenamiento suele tener como consecuencia que el modelo sea muy específico y se comporte mal para nuevos elementos, especialmente si tenemos en cuenta que el conjunto de entrenamiento puede contener ruido, lo que hará que el modelo intente ajustarse a los errores, perjudicando su comportamiento global. Éste es un problema que en general presentan todas las técnicas de aprendizaje de un modelo a partir de un conjunto de datos de entrenamiento, esto es, las técnicas de aprendizaje automático, al que se conoce como «sobreajuste» *(overfitting)*.

El modo más frecuente de limitar este problema en el contexto de los árboles de decisión y conjuntos de reglas consiste en eliminar condiciones de las ramas del árbol o de las reglas, consiguiendo con estas modificaciones la obtención de modelos más generales. En el caso de los árboles de decisión, este procedimiento puede verse como un proceso de «poda» del árbol. Esto aumentará el error de clasificación sobre el conjunto de casos de entrenamiento, pero cabe esperar que lo disminuya sobre nuevos casos no usados en la construcción del árbol.

El algoritmo C4.5 permite combinar dos tipos de poda: prepoda y pospoda. El proceso de prepoda se realiza durante la construcción del árbol impidiendo la ramificación de nodos que contengan un número de ejemplos inferior a una cierta constante.

Además, se implementa también un método de pospoda del árbol ajustado inicialmente, que consiste en simplificar el árbol eliminando un subárbol (o varios) y reemplazándolo por una única hoja o por una de sus ramas (la rama del subárbol más usada), siempre y cuando esta sustitución conduzca a una tasa de error prevista más baja. Obviamente, la probabilidad del error cometido en un nodo del árbol no se puede determinar con exactitud, y la tasa de error sobre el conjunto de entrenamiento a partir del cual fue construido el árbol no proporciona una estimación apropiada del mismo. Para estimar la tasa de error, Quinlan considera que si existe una hoja que cubre N casos clasificando incorrectamente E de ellos, se puede interpretar suponiendo que nos encontramos ante una variable aleatoria que sigue una distribución binomial en la que el experimento se repite N veces obteniendo E errores. A partir de esto se estima la probabilidad de error p_e, que será la tasa de error prevista o estimada. Para ello se realiza una estimación de un intervalo de confianza para la probabilidad de error de la variable binomial y se toma como p_e el límite superior de ese intervalo (será una estimación pesimista). El nivel de confianza para dicho intervalo viene dado por 1-*CF*, siendo *CF* un parámetro suministrado por el usuario que le permite controlar la intensidad de la poda (de forma que valores más pequeños de *CF* harán que ésta sea más acusada).

Entonces, para una hoja que cubra N casos, el número de errores previstos será $N \times p_e$. De forma similar, el número de errores previstos

asociados con un subárbol será la suma de los de cada una de sus ramas y el de éstas a su vez, la suma de los de sus hojas. De este modo, un subárbol será sustituido por una hoja o una rama, es decir, será podado, cuando el número de errores previstos para éstas sea menor que para el subárbol.

1.3. REGLAS DE CLASIFICACIÓN

Aunque los árboles de decisión representan el conocimiento de manera muy sencilla, es necesario tener presente que su inteligibilidad disminuye conforme aumenta su tamaño. Un conjunto de reglas de la forma «si» (condiciones) - «entonces» (decisión) es un mecanismo de representación del conocimiento más inteligible que los árboles de decisión, puesto que cuando el problema es complejo el árbol generado es tan grande que ni siquiera tras su poda resulta sencillo comprender el modelo de clasificación completo; por ello las reglas de clasificación son una alternativa muy popular a los árboles de decisión.

El «antecedente» o conjunto de condiciones de una regla, al igual que los nodos internos de un árbol de decisión, contiene una serie de preguntas, mientras que el «consecuente» o conclusión indica la clase de las instancias cubiertas por esa regla, o quizá una distribución de probabilidad sobre las clases. Generalmente, el antecedente es una conjunción de condiciones, aunque en algunas formulaciones de reglas los antecedentes son expresiones lógicas generales en vez de simples conjunciones.

Como hemos visto previamente, los algoritmos de inducción de árboles de decisión se basan en un enfoque «divide y vencerás»; trabajan «de arriba abajo», buscando en cada nivel el atributo en base al cual realizar la partición que mejor separa las clases, y procesando recursivamente los subproblemas que resultan de una partición. De este modo, se genera un árbol de decisión, que también puede ser representado como un conjunto de reglas de manera trivial: de cada camino desde la raíz del árbol hasta una hoja se deriva una regla cuyo antecedente es una conjunción de condiciones relativas a los valores de los

atributos situados en los nodos internos del árbol y cuyo consecuente es la decisión a la que hace referencia la hoja del árbol, esto es, la clasificación realizada.

Una característica importante del algoritmo C4.5 es que permite derivar, a partir de un árbol de decisión no pospodado, un conjunto de reglas de la forma «si» (condiciones) - «entonces» (decisión).

Aunque, como acabamos de señalar, la representación de un árbol de decisión como un conjunto de reglas es trivial, puesto que de cada camino desde la raíz del árbol hasta una hoja se deriva una regla, este modo de obtener una regla por cada hoja del árbol, sin embargo, no es más que otra manera de describir el mismo y, por tanto, tales reglas seguirán siendo mutuamente excluyentes.

Lo interesante sería poder simplificar estas reglas generadas, ya que pueden contener condiciones irrelevantes en su antecedente que podrían eliminarse sin que disminuyera la precisión del clasificador, si bien entonces estas reglas más generales podrían dejar de ser mutuamente excluyentes.

Supongamos que R es una regla que cubre $Y_1 + E_1$ casos, de los cuales E_1 son errores, esto es, ejemplos mal clasificados. Si eliminamos una de las condiciones de la regla R, obtendremos una regla más general, llamémosla R^-, que además de los casos anteriores cubrirá otros $Y_2 + E_2$ casos adicionales de los cuales clasificará incorrectamente E_2, de modo que cubre en total $Y_1 + Y_2 + E_1 + E_2$ casos y clasifica incorrectamente $E_1 + E_2$.

El procedimiento implementado en el algoritmo C4.5 para decidir si simplificar o no la regla más estricta R en favor de la regla R^- más sencilla y general es similar al empleado en la pospoda del árbol:

- Para la regla más estricta R se consideran $Y_1 + E_1$ y E_1 como parámetros procedentes de una distribución binomial en la que el experimento se repite $Y_1 + E_1$ veces obteniendo E_1 errores. Se realiza una estimación de un intervalo de confianza para la probabilidad de error de dicha distribución binomial y se toma el límite superior de ese intervalo como estimación de la probabilidad de error.

- Para la regla más sencilla y general R^- se consideran $Y_1 + Y_2 + E_1 + E_2$ y $E_1 + E_2$ también procedentes de una distribución binomial, para la cual se estima un intervalo de confianza para la probabilidad de error y se toma como estimación de esa probabilidad de error el límite superior del intervalo.

Si la probabilidad de error estimada es menor o igual que para R, se acepta R^-.

El proceso anterior se realiza para todas las condiciones que integran R y así se va eliminando sucesivamente la condición que proporcione con su eliminación la menor probabilidad de error estimada de entre todas las condiciones, siempre que tal probabilidad de error sea menor o igual que la que se obtiene sin eliminar la condición.

Este procedimiento se lleva a cabo para cada una de las reglas generadas a partir del árbol ajustado inicialmente, esto es, para cada una de sus hojas.

De este modo, se obtendrá un conjunto de reglas más simples, pero, generalmente, el número de ellas no variará. Por otro lado, no será habitual llegar a conjuntos de reglas exhaustivas y mutuamente excluyentes, sino que habrá casos cubiertos por varias reglas o bien por ninguna.

Para reducir el número de reglas se procede por grupos, conteniendo cada grupo todas las reglas que cubran una misma clase. Dentro de cada uno de esos grupos se selecciona un subgrupo de reglas basándose en el principio, al que ya hemos hecho referencia, de la «longitud de descripción mínima» o principio MDL *(Minimum Description Length)* (Rissanen, 1983). Este principio es una versión formalizada del principio de economía de Occam (o «navaja de Occam») de acuerdo con el cual cuando, en igualdad de condiciones, se pretende seleccionar entre diferentes teorías o diferentes conjuntos de reglas para explicar una determinada evidencia, se debe preferir la teoría más simple.

Formalmente, el principio MDL recomienda seleccionar la teoría *h* que minimice la expresión:

$$K(h) + K(D|h),$$

donde $K(h)$ es la complejidad en bits de describir la teoría h y $K(D|h)$ es la complejidad en bits de describir la evidencia D a partir de la teoría h (lo que incluye la descripción de los ejemplos que no son cubiertos, es decir, las excepciones).

No se considera adecuada una teoría simple con muchas excepciones, pero tampoco una teoría compleja sin excepciones. La idea del principio MDL es buscar un compromiso para cubrir bastantes ejemplos y capturar así las regularidades presentes en los datos (no subajustar) sin llegar a encontrar reglas demasiado específicas (no sobreajustar).

Una vez seleccionado un subgrupo de reglas dentro de cada grupo, para resolver los conflictos que se presenten cuando existan casos cubiertos por varias reglas correspondientes a clases distintas, se ordenan los subgrupos (dentro de cada subgrupo, el orden de las reglas no importa), comenzando por aquel que proporcione el menor número de falsos positivos, es decir, el menor número de asignaciones a la clase representada por el subgrupo de casos no pertenecientes a dicha clase.

Este primer subgrupo clasificará una serie de casos. El siguiente subgrupo será aquel que proporcione el menor número de falsos positivos con los casos restantes, y así sucesivamente.

Para contemplar la situación de que existan casos no cubiertos por ninguna de las reglas así construidas, se define además una regla o clase por defecto, que será la clase mayoritaria entre todos los casos del conjunto de entrenamiento no cubiertos por ninguna regla de las que constituyen los conjuntos de reglas representativos de cada clase. En caso de empate, se resuelve a favor de la clase con mayor frecuencia absoluta, es decir, la clase mayoritaria entre todos los casos que forman el conjunto de entrenamiento.

Una vez realizado el proceso de generalización, creados los conjuntos de reglas representativos de cada clase y definida la clase por defecto, se procede a la eliminación de reglas completas, analizando para cada uno de los conjuntos de reglas generalizadas que representan cada clase si la eliminación de alguna de las reglas que lo constituyen disminuye el error producido en la clasificación y, de ser así, se elimina dicha regla, repitiéndose este proceso hasta que la eliminación de reglas ya no suponga ninguna mejora en la clasificación.

1.4. APLICACIONES EN EL CAMPO DEL ANÁLISIS DE LA SOLVENCIA

Para finalizar este capítulo, excluyendo otras posibles aplicaciones en campos distintos al de la Economía o, dentro de éste, otros temas de estudio diferentes al análisis de la solvencia empresarial, en la siguiente tabla se resumen muy brevemente los trabajos más relevantes en el ámbito académico que emplean sistemas de inducción de árboles de decisión y conjuntos de reglas.

Tabla 1.2. Estudios previos sobre el análisis de la solvencia mediante sistemas de inducción de árboles de decisión y conjuntos de reglas

Autor(es)	Técnica(s) de inducción de árboles y/o reglas	Resumen
Marais *et al.* (1984)	RPA *(Recursive Partitioning Algorithm)*	Mediante el algoritmo de particiones recursivas (RPA), presentado originalmente por Friedman (1977), y sobre el que está basado el sistema CART, se trata de modelizar las calificaciones crediticias que un banco comercial otorga a una serie de empresas, usando como variables explicativas 20 variables financieras (13 de ellas ratios) y 6 no financieras. También se emplea una técnica estadística, Probit Multinomial, alcanzándose resultados similares con ambos procedimientos, sin que pueda afirmarse que RPA sea significativamente superior.
Frydman *et al.* (1985)	RPA	Este estudio versa sobre la predicción de la quiebra. Se aplica el algoritmo de particiones recursivas y también una técnica estadística, Análisis Discriminante, partiendo de 20 variables financieras de una muestra formada por 58 empresas industriales que fueron a la quiebra durante el período 1971-1981 y 142 no quebradas seleccionadas aleatoriamente para el mismo período, escogiendo los años dentro de dicho período también de forma aleatoria, sin emparejarlos con los años exactos del grupo de las quebradas. Se concluye que los modelos RPA, además de ser más sencillos, proporcionan mejores resultados en la clasificación que las funciones discriminantes.

Autor(es)	Técnica(s) de inducción de árboles y/o reglas	Resumen
Messier y Hansen (1988)	ID3	En este trabajo se emplea el algoritmo ID3 (predecesor de C4.5), además de Análisis Discriminante, para el pronóstico de situaciones de quiebra y de incapacidad de las empresas para la devolución de préstamos bancarios. Sólo parte de los conjuntos de atributos que incluyen los modelos discriminantes coinciden con los extraídos mediante ID3, y en cuanto a resultados, ID3 se muestra claramente superior.
Tam y Kiang (1992)	ID3	Se trata de predecir el fracaso de entidades bancarias mediante la utilización del algoritmo ID3 así como de Análisis Discriminante, Regresión Logística, K-vecinos más cercanos y una red neuronal *backpropagation*. Con esta última se logran los mejores resultados. En cuanto al algoritmo ID3, éste no mejora al Análisis Discriminante ni a la Regresión Logística, pero sí al método de los K-vecinos más próximos.
McKee (1995a)	ID3	En este trabajo se intenta predecir el fracaso empresarial empleando una muestra formada por 60 sociedades cotizadas en bolsa. Mediante el algoritmo ID3 se derivan varios árboles de decisión que se podan según el criterio del investigador para formar un modelo final consistente con una teoría acerca de la continuidad empresarial propuesta previamente por el autor. El modelo consta de 2 ratios financieros y 2 reglas, y alcanza una precisión del 97 % en la clasificación.
McKee (1995b)	ID3	Este estudio es una extensión del anterior. En primer lugar, se aplica el modelo original a otra muestra de sociedades cotizadas en bolsa formada esta vez por 202 empresas (mitad sanas y mitad fracasadas) y, dado que la precisión en la clasificación es globalmente más reducida (65 %), se plantea la cuestión de si el cambio en el rendimiento del modelo se debe a las variables empleadas o a los valores umbrales seleccionados para las mismas. Entonces, se utiliza el algoritmo ID3 para recalcular

Autor(es)	Técnica(s) de inducción de árboles y/o reglas	Resumen
		los valores umbrales óptimos para la muestra de 202 compañías, obteniéndose un nuevo modelo con los mismos dos ratios pero diferentes umbrales, que clasifica correctamente el 92 %, 82 % y 79 % de las empresas en uno, dos y tres años antes del fracaso, respectivamente. Sobre la muestra original de 60 compañías, este nuevo modelo alcanza una precisión del 87 %.
Bonsón Ponte *et al*. (1996)	*XpertRule*	Mediante el sistema de inducción *XpertRule* de Attar Software Limited, que se apoya en el algoritmo ID3, se analiza la crisis del sector bancario español, empleando el mismo conjunto de ejemplos que Serrano Cinca y Martín del Brío (1993). Esta base de datos consta de 66 bancos españoles del período 1977-1985, 29 de ellos en situación de quiebra, caracterizados por 9 ratios financieros. Para obtener una medida válida de la capacidad predictiva del árbol de decisión, se utilizó la técnica *jackknife,* obteniéndose un porcentaje de acierto del 77,3 %. Para mejorar los resultados, se aplicó la denominada por los autores «técnica del árbol de moda», consistente en, partiendo de los resultados anteriores, emplear el método *jackknife* pero forzando que todos los árboles tuviesen como base la combinación de ratios que se repetía un mayor número de veces. De este modo, se clasificó correctamente el 83,33 % de los casos. Si bien el porcentaje de acierto del modelo inducido fue menor que el obtenido en el trabajo de referencia mediante el empleo de redes neuronales artificiales (94,45 %), mostrando, por tanto, una menor capacidad predictiva, la capacidad explicativa del árbol es indudable, revelándose como determinantes de la situación de quiebra variables de liquidez, rentabilidad financiera y rentabilidad económica.

Autor(es)	Técnica(s) de inducción de árboles y/o reglas	Resumen
Jeng *et al*. (1997)	ID3 FILM *(Fuzzy Inductive Learning Method)*	En este trabajo, que versa sobre la predicción de la quiebra, se presenta un método de aprendizaje inductivo borroso que *integra la teoría* de los conjuntos borrosos *(fuzzy set theory)* dentro de los procesos del aprendizaje inductivo normal. El método convierte un árbol de decisión inducido por el método habitual en un árbol de decisión borroso, en el cual los umbrales de decisión y las clases asociadas con las hojas son borrosos. Los resultados empíricos indican que este nuevo enfoque *fuzzy* supera ligeramente en capacidad predictiva a ID3, y que tanto ID3 como FILM superan al Análisis Discriminante.
Dizdarevic *et al.* (1999)	CART CN2	En este estudio se aborda el problema de la predicción de la quiebra sobre una muestra de 120 empresas españolas utilizando como variables predictoras ratios financieros. Los métodos empleados son el sistema inductor de árboles de decisión CART y el inductor de reglas CN2, además de Análisis Discriminante, Regresión Logística y Redes Bayesianas. También se implementan dos técnicas procedentes de la Inteligencia Artificial para combinar los clasificadores y así mejorar la predicción de los modelos individuales: el Principio de Voto por la Mayoría y el Formalismo Bayesiano. Los resultados dependen de la técnica de validación empleada; cuando se acude a técnicas de validación cruzada, la eficiencia de los métodos de inducción y las Redes Bayesianas disminuye apreciablemente, lo que sugiere un problema de sobreajuste, especialmente para las Redes Bayesianas y el sistema de inducción de reglas CN2. De acuerdo con los autores, este problema podría ser evitado mediante la aplicación de criterios que penalicen estructuras complejas durante el proceso de aprendizaje.

Autor(es)	Técnica(s) de inducción de árboles y/o reglas	Resumen
González Pérez *et al*. (1999)	See5	Estos autores estudian el problema de la insolvencia empresarial a partir de los datos de PYMES que depositan cuentas en el Registro Mercantil de Santa Cruz de Tenerife. Para ello utilizan el sistema de inducción de árboles de decisión y conjuntos de reglas denominado See5, que se trata de la nueva versión comercial para sistemas operativos Windows del algoritmo C4.5. Los resultados que obtienen coinciden, en líneas generales, con los de trabajos anteriores sobre el mismo tema.
Martín Zamora (1999)	XpertRule	Esta investigación se realizó sobre el sector del crédito cooperativo. Al igual que en el trabajo de Bonsón Ponte *et* al. (1996), se utiliza el *software* XpertRule, pero tratándose ahora de explicar el comportamiento de las cajas rurales provinciales andaluzas ante la crisis sufrida en el período 1978-1985. Los resultados dependen de la técnica de validación empleada, utilizándose tanto el método *jackknife* como una muestra externa.
Gentry *et al*. (2002)	C4.5	En este trabajo se trata de predecir la quiebra usando como variables explicativas 12 componentes de flujos de caja. Requiriéndose para el cálculo de información sobre flujos de caja los balances y cuentas de resultados de los dos años previos a la declaración de quiebra, la muestra de empresas utilizada para el análisis consta de 198 firmas industriales (99 fracasadas y 99 no fracasadas) del período 1971-1987, emparejadas por sector de actividad y volumen de ventas. Se emplea el algoritmo C4.5 para inducir un conjunto de árboles de decisión mediante el procedimiento *jackknife*, obteniéndose una precisión media del 86 %, y se desarrolla un árbol global compuesto por atributos que aparecen al menos en el 50 % de los árboles anteriores, que clasifica correctamente el 89 % de las compañías y consta de sólo 3 atributos. Di-

Autor(es)	Técnica(s) de inducción de árboles y/o reglas	Resumen
		cho árbol global muestra que las compañías fracasadas no pagan dividendos, no invierten en nueva planta y equipo y no generan flujos de caja netos positivos procedentes de las operaciones, justo al contrario que las empresas no fracasadas. Por último, usando los mismos datos, se aplica también una técnica estadística, un modelo Probit, que alcanza una precisión del 67,5 %.
CIELEN *et al*. (2004)	C5.0	Este estudio versa sobre la predicción de la quiebra comparando un método de programación lineal, un modelo DEA *(Data Envelopment Analysis)* y el sistema de inducción de árboles de decisión y reglas C5.0, la nueva versión comercial para sistemas operativos Unix del algoritmo C4.5. Se calculan 11 ratios a partir de las cuentas anuales de las empresas que componen la muestra: 276 compañías no fracasadas y 90 fracasadas, entendiendo por tales aquellas que fueron declaradas en quiebra en Bélgica durante los años 1994, 1995 y 1996. No se incluyen empresas del sector financiero porque sus características difieren en gran medida de las de otros sectores. Los porcentajes de acierto obtenidos mediante *cross-validation* son 85,1 %, 79,9 % y 78 % para el modelo DEA, C5.0 y el modelo de programación lineal, respectivamente.

Capítulo 2:
Selección de datos y variables

2.1. SELECCIÓN DE LA MUESTRA DE EMPRESAS

La muestra de empresas que hemos utilizado en nuestro análisis es la seleccionada por Sanchis Arellano (2000) para la aplicación del Análisis Discriminante a la predicción de la insolvencia en empresas españolas de seguros no vida. Dicha muestra también fue empleada posteriormente para la predicción de crisis empresariales en seguros no vida mediante la aplicación de la metodología *rough Set* (Segovia Vargas, 2003).

Aquí nos referiremos simplemente a los aspectos más importantes acerca de la selección de la muestra. Una descripción más detallada puede encontrarse en Sanchis Arellano (2000).

La muestra abarca datos del período comprendido entre 1983 y 1993, extraídos de la publicación anual *Balances y cuentas. Seguros privados,* de la Dirección General de Seguros y Fondos de Pensiones. Consta de dos submuestras del mismo tamaño, una integrada por 36 empresas fracasadas —entendiendo por tal aquellas que fueron intervenidas por la Comisión Liquidadora de Entidades Aseguradoras (CLEA)— y la otra por 36 empresas no fracasadas —que para los mismos períodos se mantenían en funcionamiento—.

La muestra comprende, entonces, un total de 72 empresas españolas de seguros no vida[1], todas ellas sociedades anónimas, por entender

1. Las empresas seleccionadas operan fundamentalmente en los ramos de no vida, de modo que su participación en el negocio de vida será inexistente o bien tendrá un carácter minoritario o residual.

que otro tipo de formas societarias, como las mutuas o las cooperativas, poseen problemáticas demasiado diferentes para estudiarlas de modo conjunto.

El tipo de muestreo llevado a cabo fue por emparejamiento, siguiendo la metodología empleada por otros autores en aplicaciones similares del Análisis Discriminante: Altman (1968); Deakin (1972); Altman *et al.* (1977); López Herrera *et al.* (1994); y Martínez de Lejarza Esparducer (1999).

Dado que en el conjunto de la población las empresas insolventes representan un porcentaje relativamente pequeño, interesaba contar con el mayor número posible de empresas en esa situación presentes en la población a fin de tener un conjunto estadísticamente relevante.

Una vez seleccionadas las empresas insolventes, se procedió a la selección de las empresas sanas, grupo este mucho más numeroso y del que sólo se utilizó una muestra integrada por una fracción reducida del conjunto de empresas existentes, tratando de que ambas submuestras no fuesen demasiado heterogéneas en aquellos factores que no eran objeto del estudio.

Así, cada una de las 36 empresas fracasadas se emparejó con otra no fracasada de características similares para intentar, mediante la utilización de este muestreo por parejas, que los resultados de la clasificación se debieran a la situación financiera de las empresas y no a otro tipo de factores como el tamaño de las mismas, el tipo de negocio o el período en el que la empresa desarrollase su actividad.

La variable fundamental que se utilizó para emparejar fue el año de procedencia de los datos, comprobando además que no se diese el caso de encontrarse con una submuestra de empresas sanas en la que el resto de factores mencionados se distribuyese de forma muy diferente a la de las empresas fracasadas, ya que esto podría oscurecer el papel de las variables de carácter financiero en la explicación de la insolvencia y, con ello, dificultar la interpretación de los resultados.

Para asociar los datos en función del tamaño de las empresas —medido a través del volumen de primas— y el tipo de negocio, evitando con este modo de control del experimento la influencia de dichos factores en el análisis, se utilizó la publicación anual *Estadística de Seguros Privados* elaborada por la Unión Española de Entidades Aseguradoras y Reaseguradoras (UNESPA), la organización patronal de las empresas de seguros que operan en el mercado español.

Una vez tomada la muestra, nos situamos en períodos anteriores al de la insolvencia para tratar de determinar qué indicios de este suceso nos proporcionan los datos de las cuentas anuales en forma de ratios.

El éxito o fracaso de una empresa será entendido entonces como una variable dependiente que deberá ser explicada por un conjunto de ratios financieros que actuarán como variables independientes. Hemos seleccionado un total de 25 ratios financieros, unos populares en la literatura contable para medir la solvencia empresarial y otros específicos del sector asegurador, que comentaremos más adelante.

Al objeto de comprobar el poder explicativo de los ratios en diferentes horizontes temporales hemos tomado de cada empresa fracasada interviniente en la muestra las cuentas anuales de los tres años previos a la quiebra, considerando como año base el primer año anterior a la misma, y, dado que se ha llevado a cabo el muestreo por emparejamiento mencionado, tomaremos también para su pareja las cuentas anuales de tres años consecutivos partiendo del año base.

De este modo, desarrollaremos diferentes modelos según que los datos procedan del primer, segundo o tercer año previo a la quiebra, tratando así de predecir la crisis con uno, dos o tres años de antelación, respectivamente.

Las empresas utilizadas en el estudio se recogen en las siguientes tablas (tabla 2.1 y tabla 2.2).

Tabla 2.1. EMPRESAS INTERVENIDAS POR LA CLEA

N.º	NOMBRE	CÓDIGO	AÑO BASE
1	Kairos, Cía. de Seguros y Reaseguros, S. A.	C-043	1993
2	Igualatorio Médico Palentino de Seguros, S. A.	C-130	1993
3	Asistencia Sanitaria 2000, S. A. de Seguros[1]	C-454	1993
4	Sociedad Andaluza de Seguros, S. A.	C-507	1993
5	Consegur, S. A. de Seguros Generales	C-598	1993
6	Unión Social de Seguros, S. A.	C-638	1992
7	Mundi-Seguros, Cía. de Seguros y Reaseguros, S. A.	C-663	1992
8	Apolo, Compañía Anónima de Seguros	C-008	1991
9	Unión Europea de Seguros, S. A.	C-568	1991
10	Segurauto, S. A. Cía. de Seguros y Reaseguros	C-573	1991
11	Reunión Grupo 86 de Seguros y Reaseguros, S. A.[2]	C-440	1990
12	Servicios Médicos, Cía. de Seguros, S. A.	C-450	1990
13	Larra, S. A. de Seguros Generales	C-561	1990
14	Unión Alicantina de Seguros, S. A.	C-567	1990
15	Mades Fondo Asegurador, S. A. de Seguros	C-664	1990
16	Técnica Aseguradora de Seguros y Reaseguros, S. A.	C-352	1990
17	Unión Peninsular de Seguros, S. A.	C-555	1990
18	Mas Grupo 86 Salud de Seguros y Reaseguros, S. A.[3]	C-581	1989
19	Munauto, S. A. Seguros	C-608	1989
20	Unión Ibérica Grupo 86 de Seguros y Reaseguros, S. A.[4]	C-523	1989
21	Sociedad Occidental de Seguros, S. A.[5]	C-615	1989
22	España Vitalicia, S. A.	C-071	1989
23	Instituto Médico Quirúrgico, S. A.	C-422	1988
24	Madrid, S. A. de Seguros Generales	C-111	1986
25	Compañía Mercantil de Seguros, S. A.	C-560	1986
26	Médica Riojana, S. A.	C-460	1986
27	Igualatorio Vallisoletano Médico-Quirúr. y de Espec., S. A.	C-328	1986
28	Igualatorio Médico Nuestra Señora del Rosario, S. A.	C-321	1986
29	Argüelles, S. A. de Seguros	C-289	1986
30	Palace, S. A.	C-250	1985
31	Asociación Clínica Española, S. A.	C-283	1984
32	Cosmos, Cía. de Seguros Generales, S. A.	C-564	1984
33	Alianza Previsora, S. A.	C-220	1984
34	Clínica Argüeso, S. A.	C-224	1984
35	Labor, S. A.	C-330	1984
36	La Gloria Eterna, S. A.	C-474	1984

Notas:

1. Hasta 1991 Policlínica Santiago, S. A.
2. Hasta 1987 Reunión de Seguros y Reaseguros, S. A.
3. Hasta 1987 Mas, S. A. (Seguros y Reaseguros).
4. Hasta 1987 Unión Ibérica de Seguros y Reaseguros, S. A.
5. Sólo operó durante 2 años.

Tabla 2.2. EMPRESAS SANAS

N.º	NOMBRE	CÓDIGO	AÑO BASE
101	Metrópolis, S. A. Cía. Nacional de Seguros y Reaseguros	C-121	1993
102	Igualatorio Médico Leonés de Seguros, S. A.	C-403	1993
103	Alergia, Cía. de Seguros de Asistencia Sanitaria, S. A.	C-286	1993
104	Seguros Mercurio, S. A.	C-630	1993
105	Génesis Seguros Generales, S. A. de Seguros y Reaseg.	C-695	1993
106	Seguros Lagun-Aro, S. A.	C-572	1992
107	La Unión Alcoyana, S. A. de Seguros y Reaseguros	C-188	1992
108	Athena, Cía. Ibérica de Seguros y Reaseguros[1]	C-228	1991
109	Lepanto, S. A. Compañía de Seguros y Reaseguros	C-108	1991
110	Federación Ibérica de Seguros y Reaseguros, S. A.	C-076	1991
111	La Patria Hispana, S. A. de Seguros y Reaseguros	C-139	1990
112	Asociación Médica Conquense, Cía. de Seguros, S. A.	C-313	1990
113	Aseguradora Universal, S. A.	C-012	1990
114	Sur (Sociedad Anónima de Seguros y Reaseguros)	C-186	1990
115	Munat Seguros y Reaseguros, S. A.	C-665	1990
116	Europa Seguros Diversos, S. A.[2]	C-508	1990
117	Hispano Alsaciana, S. A. Seguros y Reaseguros	C-061	1990
118	Nortehispana de Seguros y Reaseguros, S. A.	C-275	1989
119	Andalucía y Fénix Agrícola, S. A. de Seguros y Reaseg.	C-004	1989
120	Compañía Astra de Seguros y Reaseguros, S. A.	C-468	1989
121	La Alianza Española, S. A. de Seguros	C-002	1989
122	La Humanitaria Seguros, S. A.	C-318	1989
123	La Boreal Médica, S. A. de Seguros	C-027	1988
124	ADEA, Compañía General de Seguros, S. A.	C-378	1986
125	ASEFA, S. A.	C-522	1986
126	Igualatorio Médico-Quirúrgico Pilarista, S. A.	C-390	1986
127	Asistencia Clínica Universitaria de Navarra, S. A.	C-325	1986
128	Sanitaria Médico Quirúrgica de Seguros, S. A.	C-515	1986
129	La Antártida, Cía. Española de Seguros, S. A.	C-506	1986
130	Compañía de Seguros La Gloria, S. A.	C-229	1985
131	Federación Médica de Seguros, S. A.	C-434	1984
132	Le Mans Seguros España, S. A.	C-552	1984
133	Clinos Sanitario, S. A.	C-226	1984
134	Salus, S. A. de Seguros	C-485	1984
135	Paraíso Universal, Cía. Española de Seguros, S. A.	C-238	1984
136	Seguro Europeo, S. A.	C-319	1984

Notas:

1. Hasta 1990 DAPA, Compañía de Seguros y Reaseguros, S. A.
2. Hasta 1988 Previsur, S. A. Compañía Española de Seguros.

2.2. DEFINICIÓN DE LA VARIABLE DEPENDIENTE: LA INTERVENCIÓN DE LA CLEA

El estudio del proceso que desemboca en la insolvencia empresarial es especialmente interesante, en la medida en que afecta negativamente a un amplio abanico de agentes económicos. El cese generalizado en el pago de los créditos afecta no sólo a una pluralidad de acreedores, sino también al propio deudor y a todos los que están de algún modo vinculados con la empresa (trabajadores, socios, entidades de crédito, etc.).

La ausencia de una teoría generalizada acerca del fenómeno del fracaso empresarial implica que no exista un acuerdo a la hora de conceptualizar el mismo.

De esta manera, el fracaso de una empresa puede corresponder a un amplio abanico de situaciones que comprenden desde la falta de rentabilidad de la firma hasta la insolvencia definitiva de la misma.

Tradicionalmente se ha asociado la idea de insolvencia a la de insuficiencia patrimonial, distinguiéndose entre insolvencia e iliquidez. Por insolvencia se ha entendido la situación de déficit patrimonial (pasivo superior al activo), mientras que la iliquidez presupone la suficiencia patrimonial acompañada de falta de medios dinerarios para atender las obligaciones.

Estas dos situaciones se corresponderían, respectivamente, con dos de las tres definiciones de fracaso más extendidas: el «fracaso jurídico o legal» y el «fracaso financiero».

- El fracaso financiero es el caso de una empresa que mantiene de forma continuada tensiones de liquidez provocadas por desajustes en su tesorería que le generan desequilibrios en su estructura económica y financiera.
- Cuando el importe de las deudas supera al de los activos totales de la empresa, esto es, cuando el neto es negativo, estaríamos ante la bancarrota o fracaso jurídico.

Una visión muy simplista de las insolvencias empresariales nos llevaría a distinguir las soluciones concursales (quiebra y suspensión de

pagos) según que la empresa se encuentre en estado de insolvencia definitiva o en situación de iliquidez. Está muy extendida esta idea de que la quiebra se corresponde con situaciones de insuficiencia patrimonial, que se sustancian a través de la liquidación, y de que la suspensión de pagos se vincula con situaciones de iliquidez, que se solventan con un convenio. Pero, en realidad, por lo que se refiere a la quiebra, ésta puede concluir también con un convenio que evite la liquidación y permita la continuación de la empresa. En relación con la suspensión de pagos, también pueden acogerse a este procedimiento empresas en situación de desbalance. Así, aunque su finalidad esencial sea la contraria, resulta que a través de la quiebra puede mantenerse la empresa y mediante la suspensión de pagos ésta puede ser liquidada.

Hasta aquí, sólo hemos hecho referencia a dos conceptos de fracaso empresarial, el financiero o iliquidez y el jurídico o insolvencia definitiva. Pero también podríamos definir como tal el «fracaso económico», que sería el de una empresa que no es rentable, y, por supuesto, también podrían ser equiparadas a estados de fracaso empresarial otras situaciones muy diversas que pueden darse dentro de la empresa, como, por ejemplo, la pérdida de cuota de mercado.

Obviamente, los distintos niveles de fracaso empresarial desembocarán en diferentes situaciones, como aplazamientos de deudas, reestructuración empresarial o cierre definitivo del negocio. Asimismo, empresas inmersas en situaciones similares también pueden tomar caminos muy diferentes. Así por ejemplo, como ya hemos mencionado, una empresa declarada en suspensión de pagos podría terminar liquidándose o bien conseguir un acuerdo que le permita continuar su actividad.

Como consecuencia de lo expuesto, en la amplia variedad de estudios sobre predicción del fracaso empresarial existen diferentes formas de entender el mismo como variable dependiente. En este sentido, señala Gabás Trigo (1997) que «esta diversidad de situaciones obliga a los investigadores de la insolvencia o del fracaso empresarial a definir su concepto propio de forma explícita, por lo que se utilizan variadas definiciones en función de los objetivos o en razón a la disponibilidad de datos».

Dentro de la diversidad existente de definiciones de fracaso empresarial, la más extendida es la basada en situaciones concursales, debido

a que se trata de un concepto riguroso donde no caben diferentes interpretaciones. Además, es un concepto presente en bases de datos asequibles. Ésta es una cuestión importante, al menos en nuestro país. Resulta mucho más factible formar la submuestra de empresas fracasadas cuando el fracaso se define de esta manera.

En España, a finales de los años setenta, tras algunos procesos concursales sonados, como la quiebra de la Caja Popular de Crédito y Ahorro de Cataluña o la suspensión de pagos y posterior quiebra del Banco de los Pirineos, se advirtió la necesidad de garantizar los intereses de los impositores y del propio sistema financiero, arbitrando procedimientos que garantizaran en caso de crisis la devolución —aunque limitada— de las imposiciones. Para este fin se constituyeron los Fondos de Garantía de Depósitos (en bancos, cajas de ahorro y cooperativas de crédito). Además, se posibilitó la intervención de las entidades de crédito por el Banco de España. Así, por ejemplo, esta intervención se llevó a cabo en el conocido caso de Banesto, a pesar de reunir los presupuestos que hubieran podido desencadenar su quiebra.

Otro sector en el que las crisis empresariales alcanzan a una multitud de economías es el asegurador. Cuando en los años ochenta la crisis afectó a las entidades aseguradoras (principalmente las de pequeña dimensión) se instrumentó un procedimiento que permitiera garantizar —también con alcance limitado— la cobertura de los riesgos asumidos por las empresas aseguradoras insolventes. A esta finalidad responde el Consorcio de Compensación de Seguros. Al mismo tiempo se fue conformando lo que se conoce como la CLEA (Comisión Liquidadora de Entidades Aseguradoras), a la cual se podía encomendar la liquidación de la compañía, sin que en estos casos fuese preceptivo acudir a la suspensión de pagos o a la quiebra.

En consecuencia, en nuestro país, la definición de «fracaso empresarial» puede ser muy precisa cuando se analizan sectores concretos de la economía como el bancario y el del seguro. Así, algunos autores españoles definen la variable dependiente o fracaso empresarial en sus modelos de predicción de crisis del sector bancario como la intervención del Fondo de Garantía de Depósitos (Laffarga Briones *et al.*, 1985; Pina Martínez, 1989). En el sector del seguro, también algunos autores,

por ejemplo, Mora Enguídanos (1994), definen el concepto de «fracaso empresarial» como la intervención de la CLEA.

Como hemos mencionado anteriormente, ésta es la definición de variable dependiente que se ha considerado aquí, debido a la objetividad de dicha medida de fracaso. Entenderemos entonces por empresa fracasada aquella que ha tenido que ser intervenida por la CLEA. En palabras de Sanchis Arellano *et al*. (2003) «con esto nos aseguramos el estar trabajando realmente con una muestra de empresas que han desaparecido por problemas financieros permanentes, evitando trabajar con empresas con problemas temporales o que se han liquidado de forma voluntaria».

La particular crisis del sector asegurador a la que nos hemos referido hizo necesario adoptar una serie de medidas de carácter urgente que permitieran la liquidación ordenada y ágil de las entidades aseguradoras cuya liquidación fuera intervenida administrativamente. Es en el marco de estas medidas cuando se crea la Comisión Liquidadora de Entidades Aseguradoras, cuyo objeto era asumir, en determinados casos, la condición de liquidador en el supuesto de entidades intervenidas.

Aunque la CLEA en su creación a través de un Real Decreto-Ley en 1984 (Real Decreto-Ley 10/1984) se concibió como un ente nacido para una situación concreta y determinada derivada de la necesidad de poner fin a problemas urgentes planteados en aquel momento cuya reiteración no se esperaba, lo cierto es que este organismo se ha mantenido a lo largo de casi todo este tiempo afianzándose en la normativa reguladora del seguro. Así, la Ley de Ordenación y Supervisión de los Seguros Privados de 1995 (Ley 30/1995) se ocupaba detenidamente de la regulación de la entidad, su definición, objeto y funciones y los procedimientos que había de seguir en el desarrollo de sus tareas propias, confirmando su consolidación como una pieza clave para la ordenación y saneamiento del sector del seguro español.

Como señala Sanchis Arellano (2000), «la CLEA ha sido una experiencia positiva, ya que representa una respuesta adecuada y propia de una economía de mercado para la crisis y el saneamiento del sector asegurador español».

Durante el período abarcado por la muestra de empresas que hemos utilizado para el desarrollo de nuestro estudio empírico, la CLEA, organismo autónomo de la Administración General del Estado, con personalidad jurídica propia y plena capacidad de obrar para el cumplimiento de sus funciones, vinculada a la Administración General del Estado a través del Ministerio de Economía y Hacienda, ejerciente del control de su eficacia mediante la Dirección General de Seguros y Fondos de Pensiones, tenía por objeto asumir la liquidación de aquellas entidades aseguradoras que por encontrarse en una situación patrimonial irregular, de previsible insolvencia, le fuese encomendada por el Ministro de Economía y Hacienda o por el órgano competente de la respectiva Comunidad Autónoma.

En la actualidad, y por disposición de la Ley de Medidas de Reforma del Sistema Financiero de 2002 (Ley 44/2002), la actividad de la CLEA ha pasado a ser desempeñada por el Consorcio de Compensación de Seguros (CCS).

El CCS es una entidad pública empresarial, adscrita al Ministerio de Economía y Hacienda a través de la Dirección General de Seguros y Fondos de Pensiones, con personalidad jurídica propia y plena capacidad de obrar. En su actividad, la entidad está sujeta al ordenamiento jurídico privado, lo que significa que el Consorcio ha de someterse en su actuación, al igual que el resto de las entidades de seguros privadas, al Texto Refundido de la Ley de Ordenación y Supervisión de los Seguros Privados (Real Decreto Legislativo 6/2004) y a la Ley de Contrato de Seguro (Ley 50/1980).

Desde sus orígenes, que se remontan a mediados del siglo pasado, el CCS ha estado al servicio del sector asegurador complementando las coberturas de éste en la atención de determinadas necesidades sociales especialmente difíciles de asumir por el mercado, basándose en los principios de solidaridad, compensación, colaboración y subsidiariedad. Durante su trayectoria el Consorcio ha ido abarcando distintos ámbitos del seguro, en función de las mencionadas necesidades, y entre los que cabe citar las funciones que tiene encomendadas en la cobertura de los riesgos extraordinarios (catástrofes naturales y actos de grave incidencia social como el terrorismo); en el sistema del seguro agrario

combinado; en el seguro de automóviles de suscripción obligatoria; en la liquidación de entidades aseguradoras; y en otros terrenos de menor incidencia en cuanto a cúmulo de actividad, como el seguro de crédito a la exportación, el seguro obligatorio de viajeros, el seguro obligatorio del cazador y el seguro de riesgos nucleares[2].

Su Estatuto Legal fue aprobado en 1990 y, tras sucesivas modificaciones, ha quedado recogido en el Texto Refundido del Estatuto Legal del Consorcio de Compensación de Seguros (Real Decreto Legislativo 7/2004).

Las actividades del CCS se enmarcan en las funciones aseguradoras y no aseguradoras que tiene legalmente encomendadas. Respecto de las primeras cabe destacar su carácter subsidiario, siendo su actuación, por lo general, la de un asegurador directo, en defecto de participación del mercado privado, y también la propia de un Fondo de Garantía, cuando se dan determinadas circunstancias de falta de seguro, insolvencia del asegurador, etc.

Entre las funciones no aseguradoras de la entidad, se incluyen las que anteriormente se encomendaban a la CLEA, que, como ya hemos mencionado, consisten fundamentalmente en asumir la liquidación de entidades aseguradoras cuando le sean encomendadas por el Ministro de Economía y Hacienda o, en su caso, por el órgano competente de la respectiva autoridad autonómica. También actúa como interventor único en los procedimientos de suspensión de pagos y como comisario, síndico y depositario en los de quiebra[3].

Para finalizar las reflexiones acerca de la definición de la variable dependiente que hemos considerado, esto es, la intervención de la entidad por parte de la CLEA, debemos señalar que, a pesar de sus ventajas, esta conceptualización del fracaso empresarial también presenta algún inconveniente. En cuanto a la submuestra de empresas sanas, no puede asegurarse que una compañía que no haya sido intervenida por la CLEA no esté realmente atravesando dificultades financieras. Pudiera darse el caso de que una empresa en una situación financiera difí-

2. http://portal.minhac.es/Minhac/Temas/Economia/Seguros+y+Fondos+de+Pensiones/default.htm.
3. http://www.consorseguros.es/.

cil fuese absorbida en vez de liquidada. No obstante, para evitar que compañías con problemas formasen parte de la submuestra de empresas sanas se comprobó que estas empresas seguían en funcionamiento en los años posteriores a los del período muestral (SANCHIS ARELLANO *et al*., 2003).

2.3. LAS VARIABLES INDEPENDIENTES: LOS RATIOS FINANCIEROS

Durante las últimas tres décadas se ha venido desarrollando una amplia literatura empírica acerca de la predicción del fracaso empresarial basándose en datos extraídos de los estados financieros. Este tipo de investigación se originó en Estados Unidos, donde mayoritariamente se desarrolló durante los años setenta y ochenta. Como era de esperar, el origen en España de la investigación en torno a la predicción del fracaso empresarial fue mucho más tardío, en la década de los ochenta, y alcanzó un auge importante en la pasada década y en la actual.

El análisis de los estados contables de una empresa facilita la emisión de un diagnóstico sobre la misma, en la medida en que dichos estados ponen de manifiesto en forma cuantitativa, objetiva y sistemática la realidad económica y financiera de la empresa.

Sin duda el análisis de la información contable de una entidad se ha instrumentado principalmente a través del cálculo de ratios. Este análisis mediante ratios de la información contenida en los estados financieros es indiscutiblemente uno de los más útiles para determinar la situación económico-financiera de una empresa, sobre todo desde la perspectiva externa.

Los ratios permiten sintetizar y comparar gran cantidad de información contable que de otra manera sería difícil de interpretar.

La utilidad de los ratios y, con ella, su popularidad, descansa en su simplicidad y al mismo tiempo su capacidad de expresar información complementaria y distinta a la proporcionada por las magnitudes que se ponen en relación. Un ratio aporta un valor añadido a los componentes del mismo analizados individualmente.

Esta expresión simple permite además realizar comparaciones entre diferentes empresas y distintos períodos de tiempo, posibilitando el establecimiento de conclusiones acerca de los cambios que se producen en su valor.

No obstante, a la hora de efectuar la interpretación de los ratios debemos tener en cuenta también alguna de sus limitaciones, como lo es su carácter estático. Aun cuando incluyan alguna magnitud dinámica, los ratios representan una información en un momento del tiempo, definen la situación de la empresa en relación con una magnitud determinada en aquel momento concreto del tiempo al que se refiere la información.

Por otro lado, siendo una de sus principales ventajas el permitir establecer comparaciones en el tiempo y en el espacio, así como diferencias en la normativa, en los usos y en las prácticas contables dificultan dicha comparación.

Al proceder al análisis financiero de una entidad, hemos de considerar los ratios como indicadores que habrá que analizar como un conjunto. Generalmente, un solo ratio es insuficiente, e incluso equívoco, para realizar juicios de valor. Tampoco es conveniente calcular una gran cantidad de ratios similares y tratar de formarse un juicio a partir de ellos; por contra, es más apropiado seleccionar un grupo reducido de indicadores relevantes que en conjunto definan las características económico-financieras de la entidad (Jiménez Cardoso *et al*., 2000).

Éste es el enfoque que hemos seguido aquí a la hora de seleccionar las variables más significativas para ser incluidas en nuestro modelo de predicción del fracaso empresarial. Hemos seleccionado 25 ratios financieros que, a nuestro juicio, cubren gran parte de los aspectos que interesa analizar en una empresa de seguros en relación con la solvencia, 25 ratios en principio relevantes de cara a anticipar el fracaso.

Dicha selección se ha efectuado atendiendo a los mismos criterios empleados por Beaver (1966) y la mayoría de los autores de trabajos posteriores sobre la predicción de la quiebra, a saber, ratios populares en la literatura contable para medir la solvencia de la empresa y ratios que han funcionado bien en estudios previos. Asimismo, hemos tenido en cuenta que cada sector económico posee ciertos elementos diferenciales que exigen el desarrollo de ratios específicos que sean relevantes para poder obtener una información válida. En este sentido, la mayor parte de nuestros ratios han sido específicamente propuestos para valorar la solvencia de las entidades aseguradoras.

Por otro lado, aunque los modelos de predicción del fracaso empresarial utilizan en su mayoría, al igual que aquí, ratios financieros como variables independientes, sin duda sería importante la consideración de otro tipo de variables, como macroeconómicas o cualitativas, que permitan introducir en los modelos información adicional que no se refleja en los estados contables y que además puedan paliar en alguna medida la fuerte influencia que los procedimientos contables empleados por la empresa para la elaboración de los mismos ejercen en el valor de los ratios financieros.

Posiblemente, y como ya se ha puesto de manifiesto en algunos trabajos, la inclusión conjunta en los modelos de predicción de variables financieras, macroeconómicas y cualitativas, incrementaría el poder predictivo de los mismos.

Pero, en este aspecto, nuestro trabajo se ha visto limitado por las características de la información de partida acerca de las empresas de la muestra. En concreto, tan sólo disponemos del balance de situación y la cuenta de pérdidas y ganancias de las empresas, con lo que, obviamente, resulta imposible la introducción en el modelo de otras variables distintas a las variables cuantitativas extraídas de dichos estados financieros.

Por otra parte, el período de tiempo abarcado por la muestra también ha sido un importante condicionante en la selección de las variables. A lo largo de todo el período muestral estaba vigente el Plan de Contabilidad de las Entidades Aseguradoras aprobado en 1981 (Orden de 30 de julio), que fue sustituido por el Plan de 1997 (Real Decreto 2014/1997) —que ha sido a su vez recientemente modificado[4] (Real Decreto 298/2004)— para adaptarse a la Directiva relativa a las cuentas anuales y a las cuentas consolidadas de los grupos de seguros (Directiva 91/674/CEE). Este nuevo Plan es de aplicación a los ejercicios iniciados a partir de la fecha de su aprobación, 31 de di-

4. Las modificaciones introducidas afectan a las normas de valoración referentes a las inversiones materiales, a la amortización del fondo de comercio y al cálculo de la provisión para primas pendientes de cobro en caso de fraccionamiento de las primas. Además, en concordancia con la correlativa modificación del Reglamento de Ordenación y Supervisión de los Seguros Privados (Real Decreto 297/2004), se elimina la dispensa de consolidación en el caso de entidades aseguradoras españolas dominantes que son a su vez dominadas por una entidad aseguradora domiciliada en un Estado del Espacio Económico Europeo.

BALANCE DE SITUACIÓN

ACTIVO	
I. ACCIONISTAS por desembolsos no exigidos	**A1**
II. INMOVILIZADO	**A2**
1.- Gastos de establecimiento y otros amortizables	A21
2.- Inmaterial	A22
3.- Material	A23
III. INVERSIONES	**A3**
1.- Materiales	A31
2.- Financieras	A32
3.- Inversiones en empresas del grupo, asoc. y particip. y acciones ppias.	A33
IV. PROVISIONES TÉCNICAS DEL REASEGURO CEDIDO Y RETROC.	**A4**
1.- Provisiones técnicas para riesgos en curso	A41
2.- Provisiones matemáticas (Vida)	A42
3.- Provisiones técnicas para prestaciones	A43
4.- Otras provisiones técnicas	A44
V. CRÉDITOS	**A5**
1.- Entidades y Pools de Seguros y Reaseguros	A51
2.- Créditos contra agentes	A52
3.- Provisiones (a deducir)	A53
4.- Créditos contra asegurados	A54
5.- Créditos fiscales, sociales y otros	A55
6.- Accionistas por los desembolsos exigidos	A56
7.- Dividendos activos a cuenta	A57
8.- Provisiones (a deducir)	A58
VI. CUENTA DE AJUSTE POR PERIODIFICACIÓN	**A6**
1.- Gastos anticipados e intereses pagados por anticipado	A61
2.- Otras cuentas de periodificación	A62
VII. EFECTIVO EN BANCOS Y OTROS ESTABLECIMIENTOS DE CRÉDITO, EN CAJA Y EN CHEQUES	**A7**
TOTAL ACTIVO I+II+III+IV+V+VI+VII	**AA**

PASIVO	
I. CAPITALES PROPIOS	**P1**
1.- Capital suscrito, Fondo Mutual o Fondo Permanente	P11
2.- Primas de emisión	P12
3.- Diferencias por actualizaciones del activo	P13
4.- Reservas	P14
5.- Resultados de ejercicios anteriores pendientes de aplic.	P15
6.- Resultado del ejercicio después del Impuesto	P16
7.- Minusvalías en valores negociables de renta fija	P17
II. PROVISIONES TÉCNICAS	**P2**
1.- Provisiones técnicas para riesgos en curso	P21
2.- Provisiones matemáticas (Vida)	P22
3.- Provisiones técnicas para prestaciones	P23
4.- Otras provisiones técnicas	P24
III. PROVISIONES PARA RESPONSABILIDADES Y GASTOS	**P3**
IV. DEPÓSITOS RECIBIDOS POR REASEG. CEDIDO Y RETROC.	**P4**
V. DEUDAS	**P5**
1.- Empréstitos	P51
2.- Deudas a establecimientos de crédito	P52
3.- Entidades y Pools de Seguros y Reaseguros	P53
4.- Deudas con agentes	P54
5.- Deudas con asegurados	P55
6.- Deudas condicionadas	P56
7.- Deudas a empresas del grupo	P57
8.- Deudas a empresas asociadas y participadas	P58
9.- Operaciones preparatorias o complementarias de seguros de vida, no acogidas a la Ley 8/1987, de 8 de junio	P59
10.- Cuentas fiscales, sociales y otras	P501
VI. CUENTAS DE AJUSTE POR PERIODIFICACIÓN	**P6**
TOTAL PASIVO I+II+III+IV+V+VI	**PP**

CUENTA DE PÉRDIDAS Y GANANCIAS

DEBE

	SEGURO DIRECTO	REASEG. CEDIDO MÁS RETROC. (–)	NEGOCIO NETO
I.- GASTOS TÉCNICOS			
1.- Gastos técnicos no vida:			
1.1.- Prestaciones y gastos pagados, no vida	DD1111	DR1111	D1111
1.2.- Provisiones técnicas para prestaciones, no vida:			
+Al cierre del ejercicio	DD11121	DR11121	D11121
-Al comienzo del ejercicio	DD11122	DR11122	D11122
1.3.- Otras provisiones técnicas, no vida			
+ Al cierre del ejercicio	DD11131	DR11131	D11131
- Al comienzo del ejercicio	DD11132	DR11132	D11132
2.- Gastos técnicos vida:			
2.1.- Prestaciones y gastos pagados, vida	DD1221	DR1221	D1221
2.2.- Provisiones técnicas para prestaciones, vida:			
+Al cierre del ejercicio	DD12221	DR12221	D12221
-Al comienzo del ejercicio	DD12222	DR12222	D12222
2.3.- Provisiones matemáticas, vida:			
+Al cierre del ejercicio	DD12231	DR12231	D12231
-Al comienzo del ejercicio	DD12232	DR12232	D12232
2.4.- Otras provisiones técnicas, vida:			
+Al cierre del ejercicio	DD12241	DR12241	D12241
-Al comienzo del ejercicio	DD12242	DR12242	D12242
TOTAL GASTOS TÉCNICOS VIDA Y NO VIDA	DD1	DR1	D1
II.- COMISIONES Y OTROS GASTOS DE EXPLOTACIÓN			
1.- Comisiones y participaciones:			
1.1.- Comisiones, no vida, del ejercicio	DD2111		D2111
1.2.- Comisiones, vida, del ejercicio:			
+ Comisiones y participac. de primas devengadas del año	DD21121		D21121
- Comisiones del año llevadas al activo	DD21122		D21122
+ Amort. en el año de las comisiones de adquis. llevadas al activo	DD21123		D21123

HABER

	SEGURO DIRECTO	REASEG. CEDIDO MÁS RETROC. (–)	NEGOCIO NETO
I.- PRIMAS Y RECARGOS			
1.- Primas adquiridas, no vida:			
1.1.- Primas y recargos netos de anulaciones, no vida	HD1111	HR1111	H1111
1.2.- Provisiones técnicas para riesgos en curso, no vida:			
+ Al comienzo del ejercicio	HD11121	HR11121	H11121
- Al cierre del ejercicio	HD11122	HR11122	H11122
1.3.- Provisiones para primas pendientes, no vida:			
+ Al comienzo del ejercicio	HD11131	HR11131	H11131
- Al cierre del ejercicio	HD11132	HR11132	H11132
2.- Primas adquiridas, vida:			
2.1.- Primas y recargos netos de anulaciones, vida	HD1221	HR1221	H1221
2.2.- Provisiones para primas pendientes, vida:			
+ Al comienzo del ejercicio	HD12221	HR12221	H12222
- Al cierre del ejercicio	HD12222	HR12222	H12221
TOTAL PRIMAS ADQUIRIDAS VIDA Y NO VIDA	HD1	HR1	H1
II.- OTROS INGRESOS DE EXPLOTACIÓN			
1.- Ingresos accesorios a la explotación			H21
2.- Provisiones aplicadas a su finalidad			H22
TOTAL OTROS INGRESOS DE EXPLOTACIÓN			H2
III.- INGRESOS FINANCIEROS			
1.- Ingresos de inversiones materiales			H31
2.- Ingresos de inversiones financieras			H32
3.- Ingr. de invers. en empresas del grupo, asoc. y particip.			H33
4.- Ingresos financieros varios			H34
5.- Provisiones aplicadas a su finalidad			H35
6.- Beneficios por diferencias de cambio de divisas			H36
7.- Beneficios en realización de inversiones materiales			H37
8.- Beneficios en realización de inversiones financieras			H38
TOTAL INGRESOS FINANCIEROS			H3

1.3.- Gastos de agencia	DD2113	D2113
2.- Otros gastos de explotación:		
2.1.- Sueldos y salarios		D2221
2.2.- Cargas sociales		D2222
2.3.- Dotaciones del ejercicio para amortizaciones		D2223
2.4.- Dotaciones a las provisiones		
2.5.- Gastos de explotación varios		
3.- Comisiones y participaciones del reaseguro (-):		
No vida		D231
Vida		D232
TOTAL COMISIONES Y OTROS GASTOS DE EXPLOTACIÓN		D2
III.- GASTOS FINANCIEROS		
1.- Gastos de inversiones materiales (incluidas amortizaciones)		D31
2.- Gastos inversiones financieras		D32
3.- Gastos inversiones en empresas del grupo, asociadas y participadas		D33
4.- Gastos financieros varios		D34
5.- Dotación del ejercicio para provisiones		D35
6.- Pérdidas por diferencias de cambio de divisas		D36
7.- Pérdidas en realización de inversiones materiales		D37
8.- Pérdidas en realización de inversiones financieras		D38
TOTAL GASTOS FINANCIEROS		D3
IV.- PÉRDIDAS EXCEPCIONALES		D4
TOTAL I+II+III+IV		D
V.- IMPUESTO SOBRE SOCIEDADES		D5
VI.- BENEFICIO DEL EJERCICIO DESPUÉS DEL IMPUESTO (Saldo que pasa al balance)		D6
TOTAL GENERAL		DD

IV.- BENEFICIOS EXCEPCIONALES	H4
TOTAL I+II+III+IV	H
V.- PÉRDIDA DEL EJERCICIO (Saldo que pasa al balance)	H5
TOTAL GENERAL	HH

ciembre de 1997, y, entre otros aspectos, en relación con el balance de situación se producen cambios en la ubicación de algunos epígrafes, aparición de cuentas específicas y redenominación de algunas cuentas.

También en la cuenta de pérdidas y ganancias se introduce una novedad relevante al presentarse un resultado técnico para los ramos distintos del de vida, otro para el ramo de vida y una cuenta no técnica. Haber dispuesto de esta información nos habría permitido indudablemente una mayor precisión en la definición de los conceptos manejados en los ratios.

Asimismo, en el Plan de 1997 también se establece la obligatoriedad de que se incluyan en las cuentas anuales, concretamente en la memoria, los estados de cobertura de provisiones técnicas y de margen de solvencia.

Al no haber sido posible calcular ratios en los que interviniera el margen de solvencia por no disponer de este dato, puesto que el estado de margen de solvencia no era de carácter público con anterioridad a la aprobación de esta norma, y considerando dicha medida fundamental a la hora de evaluar la solvencia en este tipo de empresas, habría sido interesante completar el estudio seleccionando una nueva muestra de sociedades anónimas de seguros no vida que abarcase datos contables desde el año 1998 en adelante, calcular nuevos ratios incluyendo aquéllos relativos al margen de solvencia y, finalmente, obtener nuevos modelos de predicción, comprobando de este modo la validez actual de los resultados obtenidos para el período muestral anterior y si el aporte de nueva información como la referente al margen de solvencia supone cambios significativos en la capacidad predictiva de los modelos obtenidos.

Desafortunadamente para nuestros fines investigadores, aunque afortunadamente para el sector y la economía, apenas se han llevado a cabo intervenciones por parte de la CLEA de entidades aseguradoras desde 1998 hasta la actualidad y aún menos de sociedades anónimas de seguros no vida, lo que ha impedido la obtención de una muestra de empresas fracasadas significativa, lo que ha imposibilitado, por tanto, la ejecución del estudio planteado.

En las páginas anteriores, se exponen los modelos de balance y cuenta de pérdidas y ganancias (vigentes hasta la aprobación del Plan de 1997) que presentan las empresas integrantes de la muestra, junto con los códigos que hemos asignado a las partidas para el posterior cálculo de ratios.

Previamente a la definición de ratios, reclasificaremos las partidas que integran el modelo de balance expuesto de acuerdo con el criterio tradicional de estructura de masas en circulante y fijo.

Así, el activo se divide en dos grandes submasas patrimoniales, el activo fijo y el activo circulante. La pertenencia al activo fijo o circulante de un elemento vendrá marcada no por la naturaleza del mismo, sino por la función que desarrolle en el seno de la empresa, lo cual determinará la distinta forma y plazo de su conversión en liquidez.

Por su parte, el pasivo, entendido como la estructura financiera, se divide a su vez en neto, pasivo fijo o exigible a largo plazo y pasivo circulante o exigible a corto plazo.

Esta reclasificación se hace necesaria porque en alguno de los ratios que definiremos posteriormente intervienen los conceptos de fijo y circulante y, como se puede observar, el modelo de balance anteriormente expuesto no plantea una clasificación entre elementos circulantes y fijos, tal como desarrolla el Plan General de Contabilidad de 1990, sino que presenta una clasificación de los elementos por naturaleza y no por vencimiento.

Para convertir este modelo de balance en un instrumento que facilite el análisis externo y permita la adecuada interpretación de resultados seguiremos la estructura de balance propuesta por Fernández Palacios y Maestro (1991). No obstante, debemos mencionar que en el sector asegurador se presentan algunas dificultades conceptuales al clasificar ciertas partidas en circulante y fijo, como las provisiones técnicas y las inversiones, y, por ello, existen otras clasificaciones del balance de las entidades aseguradoras, como las de Rodríguez Acebes (1990) o Millán Aguilar (2000), que podrían haber sido utilizadas, afectando en consecuencia a la definición de las distintas masas patrimoniales que intervienen en los ratios.

Dicha estructura de balance es la siguiente:

ACTIVO

Activo Fijo

I. Accionistas por desembolsos no exigidos
II. Inmovilizado
III. Inversiones
  1. Materiales
  2. Financieras ①
  3. Inversiones en empresas del grupo, asociadas y participadas y acciones propias

Activo Circulante

a) Realizable
III. Inversiones
  2. Financieras ②
VI. Cuenta de ajuste por periodificación

b) Exigible
V. Créditos
IV. Depósitos recibidos por reaseguro cedido y retrocedido (a deducir de los créditos) ③

c) Disponible
VII. Efectivo en bancos y otros establecimientos de crédito, en caja y en cheques

① Tradicionalmente la doctrina considera como activo fijo la cartera de valores cuando tiene una finalidad de control de otras unidades económicas o cuando tiende a mantenerse en el activo para obtener una rentabilidad, es decir, cuando no tiene una finalidad especulativa, y suele denominarse «inmovilizado financiero». Las inversiones financieras de control suelen incluirse en los balances de las entidades aseguradoras bajo el epígrafe de «inversiones en empresas del grupo, asociadas y participadas y acciones propias». Cuando las inversiones financieras coticen en mercados organizados se incluirán en el activo realizable. De hecho, aunque se trate de empresas del grupo, si los valores representativos de las mismas cotizaran en uno de estos mercados serían susceptibles de ser incluidos dentro del activo circulante.

② Que coticen en mercados organizados.

Nota: En cuanto a las inversiones financieras, a efectos de nuestro estudio, y debido a que sólo disponemos del dato del balance corres-

pondiente al epígrafe III.2 «Inversiones financieras», consideraremos todas ellas como si cotizasen en mercados organizados, es decir, como activo circulante realizable, y sólo se considerarán inmovilizado financiero las inversiones incluidas en el epígrafe III.3 «Inversiones en empresas del grupo, asociadas y participadas y acciones propias».

③ Este epígrafe figura en el pasivo del modelo de balance. Por entender que las provisiones técnicas del reaseguro cedido no tienen autonomía como elemento patrimonial, sino que surgen y desaparecen totalmente ligadas a las de seguro directo, se ha optado por no considerarlas en el activo, sino minorando las provisiones de directo y aceptado en el pasivo. Este planteamiento obliga a minorar los depósitos recibidos de los reaseguradores por razón de tales provisiones de los saldos activos con los reaseguradores.

PASIVO

Neto Patrimonial
I. Capitales propios
Pasivo Exigible a Largo Plazo
II. Provisiones técnicas
1. Provisiones técnicas para riesgos en curso
2. Provisiones matemáticas
3. Otras provisiones técnicas
IV. Provisiones técnicas del reaseguro cedido y retrocedido (a deducir) ①
1. Provisiones técnicas para riesgos en curso
2. Provisiones matemáticas
3. Otras provisiones técnicas
III. Provisiones para responsabilidades y gastos
V. Deudas ②
Pasivo Exigible a Corto Plazo
II. Provisiones técnicas
3. Provisiones técnicas para prestaciones
IV. Provisiones técnicas del reaseguro cedido y retrocedido (a deducir) ①
3. Provisiones técnicas para prestaciones
V. Deudas ③
VI. Cuentas de ajuste por periodificación

① Este epígrafe figura en el activo del modelo de balance. Conforme ha quedado expuesto más arriba, de las provisiones técnicas de seguro directo y aceptado se deducirán las correspondientes del reaseguro cedido.

② Deudas a largo plazo.

③ Deudas a corto plazo.

Nota: A efectos de nuestro estudio, y debido al desconocimiento del vencimiento de las deudas, puesto que sólo disponemos del dato del balance, consideraremos que son a largo plazo los empréstitos, las deudas con establecimientos de crédito, las deudas con empresas del grupo, las deudas con empresas asociadas y participadas y las operaciones preparatorias o complementarias de seguros de vida no acogidas a la Ley 8/1987 de 8 de junio[5], siendo el resto de subepígrafes deudas a corto plazo.

A continuación, describiremos los 25 ratios seleccionados como variables independientes a introducir en los modelos, clasificados en tres grupos: ratios de equilibrio financiero, de gestión y de rentabilidad. Esta clasificación de los ratios en tres grupos tiene un valor exclusivamente descriptivo y tan sólo pretende mostrar la correspondencia entre dichos ratios y los distintos aspectos del negocio.

Ratios de equilibrio financiero

Según indica Millán Aguilar (2000), la relación básica de equilibrio entre los elementos patrimoniales de la actividad financiera de la sociedad es el denominado «ratio de equilibrio financiero» (R1):

$$R1 = \frac{Inversiones + Tesorería}{Provisiones\ técnicas + Depósitos\ recibidos}.$$

De este modo, se está considerando la necesidad derivada del negocio y la obligación legal (Real Decreto 2486/1998 y Real Decreto 297/2004) de invertir la provisión de siniestros pendientes en bienes convertibles en dinero para así poder enfrentar su pago en el momento

5. Partida análoga a la denominada «Fondos para adquisición de pensiones» en el modelo de balance con anterioridad a la aprobación de la Ley de Regulación de los Planes y Fondos de Pensiones (Ley 8/1987).

preciso, invertir las provisiones de primas no consumidas y de riesgos en curso para compensar posibles déficits técnicos con su rentabilidad financiera e invertir los depósitos recibidos por reaseguro cedido y retrocedido para remunerar a las aceptantes.

Un valor menor que la unidad en este ratio indicaría la existencia de un desequilibrio entre las inversiones y las obligaciones con los asegurados, es decir, un déficit de inversión, puesto que no todas las provisiones técnicas están invertidas en activos remunerados, lo que podría ocasionar problemas de solvencia en la entidad.

Un ratio complementario al anterior que indicaría el equilibrio de las masas patrimoniales vinculadas a la actividad aseguradora (no inversora) es el denominado «ratio de equilibrio patrimonial» (R2) (Millán Aguilar, 2000):

$$R2 = \frac{\text{Neto patrimonial}}{\text{Inmovilizado} + \text{Créditos} + \text{Ajustes periodificación (del activo)} - \text{Deudas} - \text{Provisión para riesgos y gastos} - \text{Ajustes periodificación (del pasivo)}}.$$

Un valor de este ratio menor que la unidad significa que las necesidades de financiación de la actividad aseguradora son superiores a los recursos propios y, en consecuencia, se está recurriendo a las provisiones técnicas como medio de recursos necesarios para desarrollar la actividad.

Como el déficit de financiación se cubre con provisiones técnicas, éstas no se pueden invertir en su totalidad, dando lugar a que el ratio de equilibrio financiero (R1) sea también menor que la unidad (déficit de inversión). Esta situación de desequilibrio indicaría posibles problemas de solvencia y baja capitalización de acuerdo con el volumen de actividad que la entidad desarrolla.

A través del ratio

$$R3 = \frac{\text{Activo circulante}}{\text{Pasivo circulante}}$$

se mide la liquidez o solvencia a corto plazo. Éste es el clásico indicador, con carácter general, de la «distancia a la suspensión de pagos» y, en ocasiones, también recibe esta denominación cuando se trata del sector del seguro (FERNÁNDEZ PALACIOS y MAESTRO, 1991; SANCHIS ARELLANO, 2000). Éstos y otros autores como Lozano Aragüés (1999) sostienen que este ratio debe superar el 100 %.

Sin embargo, en nuestra opinión, esta afirmación no distingue claramente a la empresa aseguradora de la empresa convencional. La necesidad de un fondo de maniobra positivo es de aplicación a la mayoría de las empresas (exceptuando algunas, claro está, como las grandes superficies comerciales). En otro tipo de empresas la falta de liquidez significaría probablemente una situación de suspensión de pagos. En la empresa aseguradora, el proceso productivo es de sentido inverso al convencional, ya que ésta cobra el importe de la prima antes de hacer frente al pago del siniestro u otra contraprestación y, por tanto, no deberían presentarse problemas por falta de liquidez, sino por no haber invertido adecuadamente los recursos procedentes de las primas, puesto que por la propia naturaleza del negocio la empresa de seguros en funcionamiento normal dispondrá de liquidez permanente. La existencia de problemas de naturaleza financiera no significaría entonces una situación de suspensión de pagos, sino directamente el aviso de una quiebra técnica, como consecuencia del mencionado ciclo inverso.

Como señala Millán Aguilar (2000), en una empresa de seguros los problemas de liquidez deben aparecer con posterioridad a los problemas económicos y no al contrario, como ocurre en otros sectores.

El ratio

$$R4 = \frac{Activo\ real}{Pasivo\ exigible}$$

es el denominado «ratio de garantía» (FERNÁNDEZ PALACIOS y MAESTRO, 1991; LOZANO ARAGÜÉS, 1999), pues representa la garantía que la empresa ofrece a terceros, ya que tiene en cuenta todos los bienes de la misma, excluyendo el activo ficticio.

Según indican Fernández Palacios y Maestro (1991), este ratio, en relación con las entidades aseguradoras, indica si una empresa, aun con déficit en el margen de solvencia o fondo de garantía, dispone de recursos suficientes para atender las obligaciones ya contraídas ante la posible suspensión de su actividad y apertura de un proceso liquidatorio. Precisamente por la exigencia del margen de solvencia una entidad en funcionamiento normal habrá de presentar en este ratio un valor apreciablemente superior a la unidad.

El ratio

$$R5 = \frac{Pasivo\ exigible}{Neto},$$

denominado «ratio de endeudamiento total», mide la intensidad de la deuda comparada con los fondos propios para deducir el nivel de influencia en la empresa por parte de terceros, o el grado de autonomía financiera de la sociedad.

Si bien en las empresas no financieras este ratio suele tender a la unidad (de manera que, si es superior, puede implicar una situación de ahogo financiero y, si es inferior, un exceso de fondos propios con la consiguiente caída de rentabilidad), en las entidades aseguradoras, según señalan Fernández Palacios y Maestro (1991), «… la tendencia al crecimiento del ratio de endeudamiento sólo debe venir limitada por el necesario aumento de su denominador, que deriva de los incrementos del margen de solvencia exigibles, a su vez, por el aumento en el volumen de negocio».

Teniendo en cuenta que la razón de ser del negocio asegurador descansa en el aumento constante del número de asegurados, permitiendo así el funcionamiento de la Ley de los Grandes Números, y que el principal componente del pasivo está constituido por provisiones técnicas, que recogen las obligaciones con los asegurados, este ratio sería también indicativo del grado de actividad de la empresa, de modo que si su valor es muy bajo podría deducirse que «los recursos aportados en forma de capitales propios no están siendo explotados en toda su amplitud» (Fernández Palacios y Maestro, 1991).

Por otro lado, la inversión del proceso productivo que caracteriza al negocio asegurador hace que no sea aplicable el supuesto de ahogo financiero. Pero, como indica Sanchis Arellano (2000), cuando el valor del ratio es elevado podríamos encontrarnos en una situación peligrosa si las provisiones técnicas no están correctamente cubiertas, puesto que el elevado volumen de negocio podría no estar respaldado con un adecuado nivel de inversiones.

El ratio

$$R6 = \frac{\text{Provisiones técnicas seguro directo}}{\text{Total primas seguro directo}},$$

denominado «ratio de cobertura» (Millán Aguilar, 2000), mide el nivel de cobertura que ofrecen las provisiones técnicas de la sociedad para hacer frente a las obligaciones contraídas por los ingresos del ejercicio, de manera que un valor alto significa un mejor nivel de provisionamiento.

Al objeto de analizar la influencia del reaseguro en la posición de solvencia de la entidad, consideraremos también este ratio neto de reaseguro:

$$R7 = \frac{\text{Provisiones técnicas negocio neto}}{\text{Total primas negocio neto}},$$

En las entidades aseguradoras un aspecto importante a tener en cuenta es la suficiencia de sus recursos propios. Aunque a través del margen de solvencia se esté considerando una suficiencia global de garantías, el nivel de fondos propios tiene su importancia, tal como hemos visto al hablar del ratio de equilibrio patrimonial.

Siguiendo a Millán Aguilar (2000), a continuación describiremos un bloque de ratios, denominados «ratios de apalancamiento», con el que se pretende medir el grado de cobertura que ofrecen los recursos propios frente a diversas situaciones de riesgo para la entidad, de manera que, como indica este autor, el nivel de apalancamiento aumenta el retorno del capital, pero también el riesgo de inestabilidad.

Ratios de exposición asegurada

$$R8 = \frac{\text{Provisiones técnicas seguro directo}}{\text{Fondos propios}}$$

Éste es un indicador del grado de apalancamiento de las obligaciones técnicas de la entidad en relación con sus recursos propios; mide el grado de absorción por la compañía de los errores de estimación en sus provisiones técnicas (Millán Aguilar, 2000; Segovia Vargas, 2003).

Con este ratio se pretende medir la relación entre recursos administrados (provisiones técnicas) y recursos propios. No sería aconsejable un valor muy elevado, esto es, un volumen de recursos administrados muy superior al de recursos propios, que situaría a la empresa en los límites de su capacidad de gestión, con menos margen de maniobra para hacer frente a problemas eventuales de equilibrio financiero.

Por la razón indicada anteriormente, expresaremos también este ratio neto de reaseguro cedido:

$$R9 = \frac{\text{Provisiones técnicas negocio neto}}{\text{Fondos propios}},$$

Ratios de apalancamiento asegurado

$$R10 = \frac{\text{Total primas seguro directo}}{\text{Fondos propios}}$$

Este ratio expresa el grado de apalancamiento del volumen de negocio de la empresa en relación con sus recursos propios; mide el grado de absorción por la compañía de los errores en el precio de sus productos.

No existe un límite básico predeterminado para el valor de este ratio. Según señala Millán Aguilar (2000), en Estados Unidos se habla de un valor máximo de 3, aunque esta cifra no está técnicamente soportada. En cualquier caso, no se aconseja que presente un valor elevado, pa-

ra no forzar los costes contractuales de la entidad. Este ratio es considerado por la compañía de *rating* A. M. Best Company en su sistema *Best's Rating System,* cuyo objetivo es evaluar la posición financiera relativa de cada asegurador frente al resto de la industria y predecir su capacidad de cumplir con sus obligaciones financieras (BEST, 1991).

Introduciremos también este ratio neto de reaseguro cedido:

$$R11 = \frac{\text{Total primas negocio neto}}{\text{Fondos propios}},$$

Un ratio similar a éste era uno de los utilizados en Estados Unidos por la National Association of Insurance Commissioners (NAIC) en su sistema IRIS *(Insurance Regulatory Information System)*, diseñado en 1973 para detectar problemas de solvencia con la suficiente rapidez como para prevenirla o mitigar, al menos, los daños causados por la misma (DEL POZO GARCÍA, 1997). De este modo, se pretendía ayudar a los supervisores de las entidades aseguradoras a la hora de seleccionar y ordenar las empresas que requiriesen una atención especial.

También el más actual sistema FAST *(Financial Analysis Tracking System)* desarrollado por la NAIC a comienzos de los años noventa incluye ratios similares a estos cuatro últimos ratios de apalancamiento (GRACE *et al.*, 1998).

Un conjunto de ratios que pueden ser considerados «ratios de solvencia en sentido estricto» (MARTÍN PEÑA *et al.*, 1999), indicadores de la misma tanto en un momento concreto del tiempo (solvencia estática) como estudiando su variación a lo largo de un período determinado (solvencia dinámica), son los siguientes:

$$R12 = \frac{\text{Gastos técnicos seguro directo}}{\text{Fondos propios}},$$

$$R13 = \frac{\text{Gastos técnicos negocio neto}}{\text{Fondos propios}},$$

$$R14 = \frac{\text{Gastos técnicos seguro directo}}{\text{Fondos propios} + \text{Provisiones técnicas}} \text{ y}$$

$$R15 = \frac{\text{Gastos técnicos negocio neto}}{\text{Fondos propios} + \text{Provisiones técnicas netas}}.$$

Estos cuatro ratios recogen en el numerador «la medida de los riesgos anuales, basándose en la valoración de los riesgos que realmente han ocurrido (siniestros del año) que se suponen registrados en la Cuenta de Pérdidas y Ganancias como Gastos Técnicos» (Martín Peña *et al*., 1999), diferenciándose entre seguro directo y neto con el fin de analizar las diferencias entre ambas posiciones y observar la repercusión en la solvencia.

En el denominador se muestra el soporte financiero de las empresas, bien a través de los capitales propios (ratios *R12* y *R13*), lo que sugiere un soporte global del riesgo independiente de las características anuales, bien a través de la suma de fondos propios y provisiones técnicas (totales o netas según el tipo de gastos técnicos con los que se esté comparando, ratios *R14* y *R15*), obteniéndose de este modo el soporte financiero real para el período analizado.

Para finalizar con este primer conjunto de ratios que hemos denominado «ratios de equilibrio financiero», no debemos olvidarnos del papel del reaseguro, una práctica común y altamente recomendable para cualquier compañía de seguros, ya que es una forma de diversificar el riesgo, pero para la que, al mismo tiempo, el regulador intenta evitar que su cantidad sea excesiva, de manera que no se comprometa la estabilidad del asegurador directo ante la posible insolvencia del reasegurador. En este sentido, el siguiente ratio, incluido también en el sistema IRIS de la NAIC, mide la dependencia de una entidad aseguradora con respecto al reaseguro:

$$R16 = \frac{\text{Comisiones sobre el reaseguro cedido}}{\text{Fondos propios}}.$$

En líneas generales, de acuerdo con el programa desarrollado por la NAIC, este ratio debería ser menor del 25 %. Un valor del ratio cercano al 25 % implicaría que los reguladores restringirían el uso del reaseguro, lo que afectaría a su vez a la capacidad de la empresa para emitir nuevas pólizas (SANCHIS ARELLANO, 2000).

Ratios de gestión

El indicador por excelencia de la eficiencia económica de un sistema de seguros es el «ratio de siniestralidad»:

$$R17 = \frac{\textit{Gastos técnicos seguro directo}}{\textit{Primas adquiridas seguro directo}}.$$

Como se puede observar, en el denominador figuran los ingresos adquiridos o periodificados, es decir, corregidos por la variación de provisiones, para, de este modo, correlacionar los ingresos que corresponden al ejercicio analizado con los siniestros del ejercicio cubiertos por los mismos.

Este ratio refleja el porcentaje de las primas que, contablemente imputadas al ejercicio, se consumen por los siniestros de la entidad.

Como señalan Millán Aguilar (2000) y Segovia Vargas (2003), los gastos técnicos en una empresa de seguros son el equivalente al coste industrial de una empresa de transformación, de ahí la importancia que presenta su análisis en el contexto de estas entidades. Estos autores indican que a través de la evolución y contrastación con el sector del ratio de siniestralidad se puede determinar la calidad de la gestión técnica. Además, mediante su cotejo con la base técnica de la entidad se puede apreciar el nivel de desviación respecto a las previsiones de coste y, en su caso, el déficit en el cálculo del precio de la póliza, permitiendo su reajuste con rapidez, siempre que no esté controlado por la Administración.

De este modo, si el ratio de siniestralidad en un determinado ramo o modalidad se encuentra por encima de lo habitual en otras empresas,

estando en línea con el previsto en las bases técnicas de la entidad, «deberá actuarse sobre los procesos de selección de riesgos (para evitar el aseguramiento de riesgos agravados a primas normales), sobre los departamentos de peritación o gestión (para eliminar posibles irregularidades que inciden en la elevación del coste de los siniestros), o sobre ambos aspectos» (FERNÁNDEZ PALACIOS y MAESTRO, 1991).

Si el ratio alcanza un valor análogo al del sector, pero supera al que se tuvo en cuenta en el momento de calcular la prima, deberá procederse al reajuste de la misma. Y si el ratio supera tanto al valor del sector, como al de las bases técnicas de la entidad, «procederá, asimismo, una elevación de sus tarifas por este concepto, sin perjuicio de otras medidas correctoras posteriores» (FERNÁNDEZ PALACIOS y MAESTRO, 1991).

La causa de una elevada siniestralidad puede encontrarse a su vez en el número de siniestros o en el coste medio de los mismos.

Podemos plantear también el «ratio de siniestralidad neta» (LOZANO ARAGÜÉS, 1999) o global, que refleja el efecto combinado de la siniestralidad del seguro directo y del reaseguro cedido:

$$R18 = \frac{\text{Gastos técnicos negocio neto}}{\text{Primas adquiridas negocio neto}}.$$

El «ratio global de gastos de gestión» es una medida global de análisis de los gastos de gestión considerado como un indicador de la calidad de la gestión de recursos:

$$R19 = \frac{\text{Gastos de gestión netos}}{\text{Total primas negocio neto}},$$

siendo «Gastos de gestión netos = Comisiones de seguro directo – Comisiones de reaseguro cedido + Gastos de agencia + Otros gastos de explotación».

Con este ratio se pretende medir el porcentaje que suponen los gastos de gestión netos (ya que incluyen las comisiones del reaseguro cedido) sobre el negocio realmente propio de la entidad (retenido o neto de las cesiones) (MILLÁN AGUILAR, 2000).

Podemos definir un indicador que sintetice, en una sola magnitud, la calidad de la gestión de la actividad aseguradora de forma global. Tal indicador, que pretende la evaluación de la gestión de la empresa de seguros, ha de integrar, por tanto, los componentes básicos del negocio asegurador: ingresos, gastos técnicos, gastos de gestión y reaseguro.

Este indicador que recoge la calidad integral de la actividad aseguradora es el denominado «ratio combinado», que, como señalan Millán Aguilar (2000) y Segovia Vargas (2003), ha sido establecido con carácter general dentro del análisis contable del sector del seguro como esa medida de evaluación de la gestión global de la actividad aseguradora, y se define como la suma de otros dos ratios (HAMPTON, 1991):

$$R20 = \text{Ratio de siniestralidad} + \text{Ratio de gastos de gestión} = \frac{\text{Gastos técnicos seguro directo}}{\text{Primas adquiridas seguro directo}} + \frac{\text{Gastos de gestión}}{\text{Total primas seguro directo}}.$$

La razón de la inclusión del ratio de siniestralidad dentro del ratio combinado es clara, ya que aquél correlaciona perfectamente los ingresos de un período con sus gastos técnicos mediante la periodificación de tales ingresos, y, como ya hemos señalado, es el indicador por excelencia de la eficiencia económica de la empresa de seguros.

En cuanto al ratio de gastos de gestión, debemos tener en cuenta que un indicador adecuado de la calidad de aplicación de los gastos de gestión debe cumplir la condición de relacionar dichos gastos con los ingresos que los han generado. En este sentido, el volumen de ingresos no periodificados es, según indica Millán Aguilar (2000), «una medida de relación directa con los componentes de los gastos de gestión más importante que los ingresos periodificados, puesto que su consumo principal (devengo de comisiones, gastos de agencia y apertura de pólizas), se efectúa en el momento de la suscripción, y no a lo largo de la vida de la póliza».

El ratio combinado así definido presenta fundamentalmente dos aspectos criticables. En primer lugar, se calcula como suma de relaciones con denominador distinto, y aunque esto a priori puede resultar extraño, en la práctica ambos denominadores suelen ser bastante similares.

El otro aspecto criticable es que no recoge en su diseño el reaseguro cedido. Para responder a esta cuestión, existen variantes del ratio combinado que incorporan el reaseguro cedido en función de distintos criterios.

Aquí utilizaremos el más usado en la industria del seguro de Estados Unidos para valorar la gestión técnica global de las empresas aseguradoras (Stewart, 1987), el «ratio combinado integrado», compuesto también por la suma de dos ratios que hemos comentado anteriormente:

$$R21 = \text{Ratio de siniestralidad neta} + \text{Ratio global de gastos de gestión} = \frac{\text{Gastos técnicos negocio neto}}{\text{Primas adquiridas negocio neto}} + \frac{\text{Gastos de gestión netos}}{\text{Total primas negocio neto}}.$$

Si el ratio combinado, en cualquiera de sus variantes, toma un valor menor que la unidad, significará que la gestión de la empresa ha sido eficiente, ya que los gastos habrán sido inferiores a los ingresos que los han generado; esto llevará aparejado un resultado técnico positivo. Al contrario ocurrirá cuando el ratio sea mayor que la unidad. Luego cabe esperar que el ratio para las empresas solventes sea menor que para las empresas con problemas.

Ratios de rentabilidad

Una medida de la adecuación del nivel de rendimiento obtenido en la cartera de inversiones viene dado por el siguiente ratio:

$$R22 = \frac{\text{Ingresos financieros}}{\text{Tesorería} + \text{Inversiones}}.$$

Éste sería un indicador de la rentabilidad obtenida por la inversión del activo disponible para tal fin.

También es uno de los ratios utilizados por la NAIC en sus sistemas IRIS y FAST. «Un rendimiento igual al rendimiento actual en los

fondos del mercado de dinero se considera apropiado en la mayoría de los casos» (Sanchis Arellano, 2000).

Como medida de la «rentabilidad financiera» o «rentabilidad de los recursos propios», una variable sin duda destacada en numerosos estudios previos en cuanto a eficiencia en la predicción del fracaso empresarial, consideraremos el ratio:

$$R23 = \frac{\text{Beneficio antes de impuestos}}{\text{Fondos propios}}$$

en lugar de «Beneficio neto / Fondos propios», puesto que el efecto fiscal podría ocasionar el acercamiento de la rentabilidad de las empresas con mayores beneficios a la de las empresas con beneficios menores o pérdidas y, de cara a la discriminación entre empresas sanas y fracasadas, no sería quizá muy oportuno el ratio neto de impuestos.

Por otro lado, de acuerdo con García Pérez de Lema *et al*. (1997), el bajo grado de capitalización característico de las empresas fracasadas hace que con frecuencia se obtengan valores extremos o resultados confusos, por ejemplo, la engañosa rentabilidad financiera positiva de las empresas con pérdidas y neto negativo. Además, al incluirse en los recursos propios valores altamente influenciados por la inflación, en particular la cifra de capital, el efecto de la antigüedad de la empresa puede provocar una distorsión de los resultados.

Por todo ello consideraremos también como indicador de la rentabilidad financiera el ratio de «rentabilidad de los recursos totales antes de impuestos»:

$$R24 = \frac{\text{Beneficio antes de impuestos}}{\text{Pasivo total}}.$$

En el mencionado trabajo de García Pérez de Lema *et al*. (1997), este ratio se presenta como el factor más discriminante del riesgo financiero.

Teniendo en cuenta que la cifra de beneficios puede ser, dentro de ciertos límites, «manipulable» por la empresa, conviene contrastar si el ratio:

$$R25 = \frac{\text{Cash-flow}}{\textit{Pasivo total}}$$

es más eficaz para la predicción de la insolvencia, como consecuencia de la menor manipulabilidad del *cash-flow* (entendido como recursos generados) en comparación con los resultados.

El «*Cash-flow* = Beneficio neto + Dotaciones a amortizaciones + Variación de provisiones» mide los recursos generados por la empresa y, por tanto, la capacidad de autofinanciación de la misma. Las partidas que la empresa suele utilizar para «adecuar» sus beneficios son la de provisiones y la de amortizaciones, ya que dispone de cierto margen para decidir qué cifra de amortizaciones y de provisiones llevar a la cuenta de resultados. Además, esta tradicional relatividad del beneficio se agrava en el sector del seguro debido a la importancia que las provisiones técnicas presentan en la determinación del mismo.

Por tanto, el *cash-flow*, en la medida en que no incluye el efecto de estas dos partidas, da una imagen más exacta de la evolución de la empresa.

Así, ya en el trabajo de Beaver (1966), pionero en aplicar un modelo predictivo basado en la información contenida en los ratios financieros, se constató que este ratio, que relaciona el *cash-flow* con las deudas totales, era el mejor predictor.

Pero a este primer trabajo le siguió un segundo (BEAVER, 1968) en el que demostraba que el ratio de «Beneficio neto / Activo total» era de una gran capacidad predictiva de las insolvencias puesto que su comportamiento era anormal durante los cinco años anteriores a la situación de crisis. Este último ratio (antes o después de impuestos) también se muestra eficiente en otros trabajos como los de, por citar algunos, Altman y Loris (1976), Laffarga Briones *et al*. (1987), Gabás Trigo (1990) y Lizarraga Dallo (1996), pero también en numerosos estudios se muestra más eficiente el anterior ratio de *cash-flow* (DEAKIN, 1972; MENSAH, 1983; GENTRY *et al*., 1985) o el de rentabilidad de los recursos propios (PINA MARTÍNEZ, 1989; LÓPEZ HERRERA *et al*., 1994; ARQUES PÉREZ, 1997).

Con esto queremos dejar patente que no existe un consenso en cuanto a la capacidad predictiva de estos tres últimos ratios y, debido a

esta dificultad de generalización, consideraremos los tres en nuestro estudio empírico.

Por último, en la siguiente tabla (tabla 2.3) se expone el listado de los 25 ratios con sus definiciones y la codificación representativa de las partidas de balance y cuenta de pérdidas y ganancias que intervienen en los mismos.

Tabla 2.3. Ratios empleados

Ratio	Definición	Códigos
R1	Inversiones + Tesorería / Provisiones técnicas + Depósitos recibidos	(A3+A7)/ (P2-A4+P4)
R2	Neto patrimonial / Inmovilizado + Créditos – + Ajustes periodificación (del activo) – Deudas – Provisión para riesgos y gastos – Ajustes periodif. (del pasivo)	P1/ (A2+A5+A6-P5-P3-P6)
R3	Activo circulante / Pasivo circulante	(A32+A5-P4+A6+A7)/ (P23-A43+P53+P54+P55+P56+ P501+P6)
R4	Activo real / Pasivo exigible	(A1+A2-A21+A3+A5-P4+A6+A7)/ (P2-A4+P3+P5+P6)
R5	Pasivo exigible / Neto	(P2-A4+P3+P5+P6)/ P1
R6	Provisiones técnicas seguro directo / Total primas seguro directo	P2/ (HD1111+HD1221)
R7	Provisiones técnicas negocio neto / Total primas negocio neto	(P2-A4)/ (H1111+H1221)
R8	Provisiones técnicas seguro directo / Fondos propios	P2/P1
R9	Provisiones técnicas negocio neto / Fondos propios	(P2-A4)/ P1
R10	Total primas seguro directo / Fondos propios	(HD1111+HD1221)/ P1
R11	Total primas negocio neto / Fondos propios	(H1111+H1221)/ P1
R12	Gastos técnicos seguro directo / Fondos propios	DD1/P1
R13	Gastos técnicos negocio neto / Fondos propios	D1/P1
R14	Gastos técnicos seguro directo / Fondos propios + Provisiones técnicas	DD1/ (P1+P2)
R15	Gastos técnicos negocio neto / Fondos propios + Provisiones técnicas netas	D1/ (P1+P2-A4)
R16	Comisiones sobre el reaseguro cedido / Fondos propios	(D231+D232)/ P1
R17	Gastos técnicos seguro directo / Primas adquiridas seguro directo	DD1/HD1

Ratio	Definición	Códigos
R18	Gastos técnicos negocio neto / Primas adquiridas negocio neto	D1/H1
R19	Gastos de gestión netos / Total primas negocio neto	D2/ (H1111+H1221)
R20	$\frac{\text{Gastos técnicos seguro directo + Gastos de gestión}}{\text{Primas adquiridas seguro directo + Total prima seguro directo}}$	$\frac{DD1}{HD1} + \frac{D2 + D231 + D232}{HD\ 1111 + HD1221}$
R21	$\frac{\text{Gastos técnicos negocio neto + Gastos de gestión netos}}{\text{Primas adquiridas negocio neto + Total prima negocio neto}}$	$\frac{D1}{H1} + \frac{D2}{H1111 + H1221}$
R22	Ingresos financieros / Tesorería + Inversiones	H3/ (A7+A3)
R23	Beneficio antes de impuestos / Fondos propios	(H-D)/ P1
R24	Beneficio antes de impuestos / Pasivo total	(H-D)/ (P1+P2-A4+P3+P5+P6)
R25	*Cash-flow* / Pasivo total	(D6-H5+D35+D2224+D2223+ D21123-D12242+D12241- D12232+D12231-D12222+ D12221-D11132+D11131- D11122+D11121-H35-H22+ H12222-H12221+H11132- H11131+H11122-H11121)/ (P1+P2-A4+P3+P5+P6)

Capítulo 3:
Árboles de decisión y reglas de clasificación aplicados a la predicción de insolvencias en empresas españolas de seguros no vida

3.1. INTRODUCCIÓN

Hemos llegado a la fase del estudio en la que procede demostrar la adecuación del algoritmo de inducción de árboles de decisión y reglas de clasificación C4.5 al problema concreto de la predicción del fracaso empresarial en las empresas españolas de seguros no vida, empleando para ello la muestra de empresas detallada en el capítulo precedente.

Como ya se ha mencionado, una vez tomada la muestra nos situamos en períodos anteriores al de la insolvencia para tratar de determinar qué indicios de este suceso nos proporcionan los datos de las cuentas anuales en forma de ratios. El éxito o fracaso de una empresa será entendido entonces como una variable dependiente que deberá ser explicada por el conjunto de 25 ratios financieros descritos en el capítulo anterior que actuarán como variables independientes, y desarrollaremos diferentes modelos según que los datos procedan del primer, segundo o tercer año previo a la quiebra, tratando así de predecir la crisis con uno, dos o tres años de antelación, respectivamente. Llamaremos a estos modelos *Modelo 1*, *Modelo 2* y *Modelo 3*.

Para desarrollar el *Modelo 1* se utilizarán las 72 empresas disponibles. Sin embargo, no disponemos de los datos de la totalidad de estas empresas para los años segundo y tercero previos a la quiebra. Al eliminar también las respectivas parejas de las empresas faltantes, contamos en total con 68 empresas para el desarrollo del *Modelo 2* y 54 empresas para el desarrollo del *Modelo 3*.

En cuanto a la verificación de la capacidad predictiva de los modelos, dado que los porcentajes de acierto sobre el propio conjunto de datos usado para su obtención no representan una medida adecuada de la validez de dichos modelos de cara a la clasificación de nuevos elementos, llevaremos a cabo el proceso de validación *jackknife*, originalmente debido a Maurice Quenouille, y que también es conocido como *leave-one-out* (EFRON, 1982).

Al disponer de pocos datos, reservar parte de ellos para el test supone utilizar todavía menos para la obtención de los modelos, lo que podría ocasionar que dichos modelos fueran de mala calidad. Además, el resultado sería demasiado dependiente del modo en el cual se hubiese realizado la partición del conjunto completo en dos subconjuntos disjuntos de entrenamiento y test. Dado que, generalmente, esta partición se efectúa de manera aleatoria, podría ocurrir que dos experimentos distintos realizados con el mismo método sobre la misma muestra obtuvieran resultados muy dispares.

Un mecanismo que permite evitar la dependencia del resultado del experimento del modo en el cual se realice la partición es el método *jackknife*. Siendo k el número de instancias que contenga el conjunto de entrenamiento (en nuestro caso, 72 para el *Modelo 1*, 68 para el *Modelo 2* y 54 para el *Modelo 3*), se elabora un modelo utilizando $k-1$ instancias y el caso restante se emplea para evaluar dicho modelo. Este procedimiento se repite k veces, utilizando siempre una instancia diferente para la evaluación del modelo. La estimación del error final se calcula como la media aritmética de los errores de los k modelos parciales.

Éste es un método muy atractivo por dos razones. En primer lugar, se utiliza la mayor cantidad posible de datos para el entrenamiento, lo que presumiblemente redundará de modo favorable en la calidad del modelo. En segundo lugar, el procedimiento es determinante, los resultados obtenidos con el mismo método sobre la misma muestra siempre serán los mismos y no dependerán del modo en el que se realice la partición de la muestra. El inconveniente vendría dado por el elevado coste computacional derivado del gran número de iteraciones que habrán de ser realizadas, con lo que para bases de datos de gran tamaño no se-

ría muy recomendable. Sin embargo, con pequeños conjuntos de datos como el nuestro, ofrece la oportunidad de conseguir la estimación más exacta que posiblemente pueda obtenerse.

Por otro lado, aunque nuestro trabajo está orientado a poner de manifiesto la utilidad de los árboles y reglas de decisión construidos con el algoritmo C4.5 de cara a predecir el fracaso empresarial en las empresas de seguros, esta tarea quedaría incompleta si no se realizase una comparación de los resultados alcanzados con los que se obtendrían aplicando al mismo problema alguna técnica alternativa y bien conocida que actuase como término de referencia con respecto al cual poder valorar de manera fundamentada la calidad real del método propuesto. Para realizar esta comparación hemos elegido una técnica procedente del área de la estadística: Regresión Logística. En la última parte del presente capítulo se expondrán los fundamentos de esta técnica y se llevará a cabo la comparación de resultados mencionada.

En lo que respecta al *software* empleado, para la obtención de los árboles de decisión hemos utilizado el paquete gratuito de minería de datos WEKA 3.4 desarrollado en la Universidad de Waikato en Nueva Zelanda (Witten y Frank, 2000) que incorpora una implementación en Java de la última versión publicada del algoritmo C4.5 (C4.5 *Release* 8), a la que los autores denominan J4.8. Pero dicha implementación sólo se refiere al algoritmo de generación de árboles de decisión, y no a su conversión en reglas, así que, para este fin, hemos empleado la última versión publicada del propio algoritmo C4.5, descargable gratuitamente en forma de código fuente en lenguaje C desde la página de Ross Quinlan (http://www.rulequest.com/Personal/). Este programa puede ser fácilmente compilado (y ejecutado posteriormente) sobre sistemas operativos tipo Unix utilizando el compilador GCC distribuido por la Free Software Foundation y que es actualmente el compilador estándar en un gran número de variantes de UNIX como LINUX, FreeBSD y otros. Asimismo, cabe señalar que existen nuevas versiones comerciales del algoritmo (denominadas C5.0 para Unix y See5 para Windows) que implementan mejoras y funcionalidades adicionales y se comercializan directamente por el propio Quinlan (Rulequest Research) o a través de paquetes de minería de datos como Clementine, aunque también

hay versiones de demostración gratuitas limitadas a bases de datos de pequeño tamaño (http://www.rulequest.com/). En cuanto al sistema de inducción de árboles de decisión, parece ser esencialmente el mismo que en C4.5. Sin embargo, la generación de reglas con las nuevas versiones es mucho más rápida y claramente se realiza de manera diferente. Como la forma concreta en que ésta se lleva a cabo no ha sido publicada, en el presente trabajo no hemos empleado las nuevas versiones del algoritmo.

En lo que a la aplicación de la Regresión Logística se refiere, hemos utilizado el *software* R 2.1.0 distribuido gratuitamente por CRAN Foundation (R Development Core Team, 2005). Este programa implementa una versión del lenguaje de tratamiento de datos S, desarrollado principalmente por J. M. Chambers desde los años ochenta hasta la actualidad, por el que fue galardonado en 1998 con el prestigioso premio para Sistemas de *Software* de la ACM (Association for Computing Machinery), y que ha cosechado una notable aceptación debido a su potencia y flexibilidad.

Por otra parte, antes de pasar a comentar los resultados, hemos de señalar que al estar utilizando como variables explicativas ratios financieros calculados a partir de los estados contables —balance y cuenta de pérdidas y ganancias— de las empresas, no existen valores desconocidos *(missing values)* para ninguna de dichas variables. No obstante, en el primer año anterior a la quiebra un 0,33 % de los valores de los ratios resulta ser infinito, por tomar el denominador valor nulo. Para el segundo año el porcentaje de valores infinito se reduce al 0,18 %, y al 0,22 % para el tercero. Ya que disponemos de una muestra relativamente pequeña, nos inclinamos por aprovechar los ejemplos que podrían ser desechados si dispusiésemos de una gran cantidad de casos. Pensemos en que estos casos para los que el valor de algún atributo es infinito pueden esconder en el resto de sus atributos información relevante de cara a la detección de patrones útiles a partir de los datos, así que teniendo en cuenta que el porcentaje de infinitos es ciertamente muy reducido, no tendría sentido eliminar esas empresas de la base de datos. Por ello, dado que, obviamente, no es posible operar con valores infinito, hemos optado por considerarlos como valores perdidos, puesto

que, como ya se ha mencionado, C4.5 implementa un procedimiento para poder trabajar con bases de datos que contengan valores perdidos. No ocurre lo mismo en el caso de la técnica estadística, que requiere que los vectores de la base de datos sean completos, así que cuando llegue el momento de la aplicación de este método habremos de tratar de alguna manera los valores que hemos considerado como desconocidos.

Los sucesivos resultados obtenidos mediante la aplicación a nuestra muestra de las técnicas mencionadas aparecen descritos a continuación.

3.2. RESULTADOS

3.2.1. Árboles de decisión

Durante el proceso de generación de los árboles de decisión hemos tratado de impedir en la medida de lo posible la construcción de árboles muy complejos y excesivamente ajustados a los datos del conjunto utilizado para dicha construcción que, en consecuencia, se comporten mal para nuevos elementos, esto es, tratamos en definitiva de evitar el problema de «sobreajuste» que en general presentan todas las técnicas de aprendizaje automático. Como ya se ha mencionado, el modo más frecuente de limitar este problema en el contexto de los árboles de decisión y conjuntos de reglas consiste en eliminar condiciones de las ramas del árbol o de las reglas, consiguiendo con estas modificaciones la obtención de modelos más generales. En el caso de los árboles de decisión, este procedimiento puede verse como un proceso de «poda» del árbol.

Partiendo de los datos del primer año anterior a la quiebra, hemos combinado los métodos de prepoda y pospoda implementados en C4.5 y detallados en el primer capítulo del presente trabajo al objeto de conseguir un modelo que, aunque no clasifique correctamente el 100 % de los elementos del conjunto de entrenamiento, es decir, las 72 empresas, manifieste un buen comportamiento (estimado mediante la validación cruzada *jackknife*) ante la clasificación de futuros casos. Así, tras sucesivas variaciones de los parámetros de aprendizaje (éstos son el mínimo número requerido de casos por hoja y el parámetro *CF* o factor de confianza que permite controlar la intensidad de la poda, de forma que

cuanto más pequeño sea el valor de dicho parámetro más acusada será la misma), hemos seleccionado como *Modelo 1* el siguiente árbol de decisión, que se ha obtenido exigiendo que cada hoja cubra al menos 3 empresas y mediante una poda muy intensa dada por un *CF* del 5 %.

FIGURA 3.1. ÁRBOL DE DECISIÓN *MODELO 1*

```
Test mode:    72-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

R3  <= 1.401888
|   R4  <= 1.463728: mala (19.27)
|   R4  > 1.463728
|   |   R5  <= 0.528132: mala (4.1)
|   |   R5  > 0.528132: buena (3.0)
R3  > 1.401888
|   R6  <= 0.049314: mala (5.63)
|   R6  > 0.049314: buena (40.0/7.0)

Number of Leaves  :     5

Size of the tree :      9
```

El árbol decisión se leería del siguiente modo:

- Si el ratio *R3* es menor o igual que 1.40 y el ratio *R4* es menor o igual que 1.46, la empresa será «mala» (fracasada).
- Si el ratio *R3* es menor o igual que 1.40 y el ratio *R4* es mayor que 1.46 y el ratio *R5* es menor o igual que 0.53, la empresa será «mala».

Y así continuaríamos descendiendo por el árbol, hasta completar la totalidad de sus hojas (recordemos que cada hoja del árbol se refiere a la decisión a tomar), concretamente 5. El tamaño del árbol, en este caso 9, hace referencia al número de nodos de que consta el mismo, tanto nodos internos como terminales u hojas. En el anexo 1 se presenta este árbol en forma gráfica.

Al final de cada una de las hojas se observan entre paréntesis unos valores *n* o *n*/*m*. *n* representa el número de empresas del conjunto de entrenamiento que verificando las condiciones que conducen a esa hoja

son asignadas a la clase representada en la misma y *m,* el número de errores cometidos, es decir, el número de empresas clasificadas incorrectamente por pertenecer a otra clase distinta a la asignada. En algunos casos aparecen valores fraccionarios por el tratamiento de los *missing values* incorporado en C4.5 que se ha comentado anteriormente.

Estudiemos ahora el significado de esta sencilla representación en forma de árbol de la regularidad presente en los datos. El hecho de que el ratio $R3 = \frac{Activo\ circulante}{Pasivo\ circulante}$ sea el atributo escogido en el nodo raíz nos confirma la importancia de la liquidez de cara a predecir el fracaso empresarial, y, lógicamente, en nuestra muestra las empresas «buenas» (sanas) presentan una mejor posición en liquidez que las empresas «malas» (fracasadas). Como se ha mencionado en el capítulo precedente, una empresa de seguros en funcionamiento normal generará liquidez de forma continuada, debido a la inversión del proceso productivo que se da en este tipo de entidades, por lo que no será habitual que aparezcan problemas de esta índole. Así que si bien una de las cuestiones más importantes para asegurar el buen funcionamiento de cualquier tipo de empresa es la necesidad de un colchón de liquidez que le permita hacer frente a sus deudas a corto plazo sin tener que recurrir a la realización de sus activos fijos y evitando así incurrir en una situación de suspensión de pagos, en el caso de la empresa de seguros dicha necesidad reviste una mayor importancia, pues debe ser capaz de atender los siniestros en el momento oportuno, y, por tanto, aunque en las empresas aseguradoras no quepa esperar problemas por falta de liquidez, la presencia de dichos problemas sería un claro síntoma de fracaso a corto plazo, en este caso, a un año vista. El punto de corte escogido para realizar la primera partición en el árbol en base al ratio *R3* es 1.40, de modo que cuando el valor de este ratio sea menor o igual que dicha cantidad pasaremos a un nodo hijo donde entra en juego el ratio *R4*, y si *R3* es mayor que 1.40 será *R6* el ratio examinado en el nodo hijo correspondiente.

En cuanto al ratio *R4,* recordemos que se trata del denominado «ratio de garantía»:

$$R4 = \frac{Activo\ real}{Pasivo\ exigible},$$

que indica la capacidad de la entidad para hacer frente a sus compromisos de pago a través de la liquidación de sus activos. Su valor habrá de ser, al menos, igual a la unidad, puesto que en caso contrario el patrimonio neto sería negativo y la empresa se encontraría en situación de quiebra técnica, de ahí que a este índice se le conozca también como «ratio de distancia a la quiebra». En nuestra muestra se observa que en las entidades que finalizan de forma continuada sus ejercicios con pérdidas el valor de este ratio se va reduciendo al ir decreciendo su activo real, ocurriendo incluso para 6 de ellas que en el ejercicio anterior a la quiebra dicho activo real es insuficiente para atender todas las obligaciones y, dado que probablemente la única posibilidad que evitaría que dichas empresas incurriesen en una situación de insolvencia sería una ampliación de capital que no llega a producirse, lo cual es obvio teniendo en cuenta que estas empresas no reparten beneficios; finalmente resultan ser intervenidas por la CLEA.

De acuerdo con nuestro árbol de decisión, si el activo circulante de la empresa no supera el 140 % del pasivo circulante y el activo real es menor o igual que el 146 % del pasivo exigible, la empresa será «mala», es decir, será insolvente en el próximo ejercicio, y tan sólo con estas dos condiciones se logra clasificar de forma correcta aproximadamente la mitad de las empresas fracasadas de la muestra. Si la segunda de las dos condiciones no se verifica, esto es, si el activo real supera el 146 % del pasivo exigible, no se alcanzará ninguna decisión respecto a la clasificación de la empresa, sino que se habrá de examinar el valor del ratio contenido en el siguiente nodo, el ratio *R5.*

$$R5 = \frac{Pasivo\ exigible}{Neto},$$

denominado «ratio de endeudamiento total», mide el grado de autonomía financiera de la entidad. Si bien en otro tipo de empresas este ratio suele tender a la unidad, la inversión del proceso productivo que caracteriza al negocio asegurador hace que no sea aplicable el supuesto de ahogo financiero. Además, dado que en las empresas de seguros el principal componente del pasivo está constituido por provi-

siones técnicas, que recogen las obligaciones con los asegurados, este ratio sería también indicativo del grado de actividad de la empresa y, de acuerdo con lo expuesto, será generalmente mayor para las empresas «buenas» que para las «malas», como se refleja en el árbol de decisión.

Volvamos ahora al nodo raíz. Ya hemos analizado qué ocurriría si el ratio *R3* fuese menor o igual que 1.40. Si, por el contrario, el activo circulante de la empresa superase el 140 % del pasivo circulante, habría que examinar el valor del ratio *R6*. Como hemos indicado en el capítulo anterior de este trabajo, el ratio $R6 = \frac{Provisiones\ técnicas\ seguro\ directo}{Total\ primas\ seguro\ directo}$, denominado «ratio de cobertura», mide el nivel de cobertura que ofrecen las provisiones técnicas de la entidad para hacer frente a las obligaciones contraídas por los ingresos del ejercicio, de manera que un valor alto significa un mejor nivel de provisionamiento. Fijémonos en que nuestro árbol nos recomienda considerar como «malas» aquellas empresas que posean un nivel de provisiones técnicas inferior al 5 % de los ingresos por primas, y ello es debido a que en este primer año anterior a la quiebra alguna de las empresas de nuestra muestra presenta un nivel de provisiones técnicas extremadamente bajo, posiblemente por causa de una infradotación continuada de las mismas de cara a maquillar el resultado.

Además de la indiscutible sencillez del árbol de decisión, que selecciona como variables de mayor poder discriminante 4 de las 25 de partida, éste obtiene un porcentaje de acierto en la clasificación del 90,28 % (65/72 empresas). No obstante, este elevado porcentaje de clasificaciones correctas obtenido sobre el propio conjunto de datos usado para la construcción del árbol no representa una medida adecuada de la validez del modelo de cara a la clasificación de nuevas empresas, puesto que, como ya hemos mencionado, podría estar sobreajustado a los datos de entrenamiento y comportarse mal para nuevos elementos. En cambio, la validación cruzada sí proporciona una estimación razonable de dicha validez. Pues bien, llevando a cabo el proceso de validación cruzada *jackknife* se obtiene un porcentaje de clasificaciones correctas de casi el 85 %, resultado más que aceptable que nos lleva a confiar en la bondad del modelo. Estos resultados se muestran a continuación.

FIGURA 3.2. RESULTADOS *JACKKNIFE* ÁRBOL *MODELO 1*

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          61       84.7222 %
Incorrectly Classified Instances        11       15.2778 %
Total Number of Instances               72

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.778     0.083       0.903     0.778     0.836   mala
  0.917     0.222       0.805     0.917     0.857   buena

=== Confusion Matrix ===

  a  b   <-- classified as
 28  8 |  a = mala
  3 33 |  b = buena
```

Como se puede observar, además del porcentaje global de acierto señalado, se detalla también la precisión por clase en forma de una serie de medidas usuales en el área del aprendizaje automático. Así, las dos primeras columnas de la sección titulada «*Detailed Accuracy By Class*» hacen referencia, respectivamente, a la tasa de verdaderos y falsos positivos: $TP\ Rate = \frac{TP}{TP + FN}$, siendo TP el número de «*true positives*» y FN el de «*false negatives*», representa el porcentaje de casos de la clase en cuestión que se clasifican como pertenecientes a dicha clase, esto es, el porcentaje de aciertos en la clase, mientras que $FP\ Rate = \frac{FP}{FP + TN}$, donde FP es el número de «*false positives*» y TN el de «*true negatives*», representa el porcentaje de casos asignados a la clase de los pertenecientes a la clase contraria. La medida «*Precision*» señala el porcentaje de los casos asignados a la clase que verdaderamente pertenecen a la misma, $Precision = \frac{TP}{TP + FP}$ y la medida «*Recall*» representa lo mismo que «*TP Rate*», esto es, el porcentaje de aciertos en la clase (se habla de «*recall*» o «*true positive rate*» dependiendo del ámbito en que se utilice esta medida). Por su parte, «*F-Measure*» es una media entre «*precision*» y «*recall*» ponderada de acuerdo con el número de casos cubiertos por cada una de estas medidas:

$$F\text{–Measure} = Precision \times \frac{TP + FP}{2TP+FP+FN} + \text{Recall} \times \frac{TP+FN}{2TP+FP+FN} = \frac{2TP}{2TP+FP+FN}$$

Finalmente, se muestra una matriz de confusión que señala en términos absolutos el tipo de errores cometidos, esto es, el número de empresas «malas» clasificadas como «buenas» y viceversa.

Tratando ahora de predecir la crisis con dos años de antelación, repetiremos el proceso de generación de árboles de decisión utilizando los datos del segundo año previo a la quiebra. De este modo, tras sucesivas combinaciones de los parámetros de aprendizaje seleccionamos finalmente como *Modelo 2* el árbol que se presenta debajo, donde se ha establecido que cada hoja cubra al menos 8 casos y, por tanto, no ha sido necesaria una poda posterior tan acusada como para el modelo anterior (*CF* del 30 % frente al 5 % de antes). El anexo 2 muestra este árbol en forma gráfica.

FIGURA 3.3. ÁRBOL DE DECISIÓN *MODELO 2*

```
Test mode:    68-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

R14  <= 0.399419
|   R6  <= 0.906477
|   |   R3  <= 3.361892: mala (11.0/3.0)
|   |   R3  > 3.361892: buena (10.0/1.0)
|   R6  > 0.906477: buena (13.0/1.0)
R14  > 0.399419: mala (34.0/10.0)

Number of Leaves  :     4

Size of the tree :      7
```

Tal como se observa en esta figura, el algoritmo selecciona ahora como variables más discriminantes los ratios *R14, R6* y *R3*. Dado que *R14* es el atributo considerado en el nodo raíz, se manifiesta de este modo la solvencia, medida a través de dicho ratio, como factor determinante a la hora de predecir el fracaso en las empresas de seguros con dos años de antelación. Este ratio, como hemos señalado en el capítulo precedente, recoge en el numerador la medida de los riesgos anuales, basándose en la valoración de los riesgos que realmente han ocurrido (siniestros del año), registrados en la cuenta de pérdidas y ganancias co-

mo Gastos Técnicos. El denominador, a través de la suma de fondos propios y provisiones técnicas, muestra el soporte financiero real de las empresas para el período analizado. De esta manera, de acuerdo con el árbol de decisión, cuando la siniestralidad supere el 40 % de la suma de fondos propios y provisiones técnicas la empresa será considerada fracasada, verificando esta condición aproximadamente el 70 % (24/34) de las empresas de la muestra que finalmente fueron intervenidas por la CLEA. Si esta condición no se cumpliese y, por tanto, la empresa no pudiese ser catalogada directamente como «mala», el siguiente ratio a examinar sería ***R6,*** con lo que de nuevo se confirma la importancia del nivel de cobertura que ofrecen las provisiones técnicas a la hora de predecir el fracaso de las empresas aseguradoras, en esta ocasión con dos años de antelación. Así, cuando la cuantía de provisiones técnicas supere el 91 % de la de ingresos por primas, la empresa será «buena». Si, por el contrario, el nivel de provisiones técnicas fuese menor o igual que el 91 % de los ingresos por primas, entraría en juego el ratio ***R3,*** manifestándose de nuevo la liquidez como factor determinante en la predicción del fracaso empresarial, aunque en esta ocasión ***R3*** no se trata del ratio considerado en el nodo raíz, sino que aparece anidado con los ratios ***R14*** y ***R6,*** de manera que si la siniestralidad no supera el 40 % de la suma de fondos propios y provisiones técnicas y además la cuantía de éstas es menor o igual que el 91 % de la de ingresos por primas, dependiendo de que la empresa alcance el elevado nivel de liquidez señalado en el árbol o no lo haga será «buena» o «mala», respectivamente.

En cuanto a la precisión clasificatoria de este modelo, cabe señalar que se trata de un modelo no sobreajustado en absoluto, ya que obtiene el mismo porcentaje de acierto sobre el conjunto de entrenamiento que el estimado a través del método *jackknife*, en concreto, el 78 %. Como cabía esperar, se observa una disminución en la precisión clasificatoria ante el aumento del horizonte temporal de la predicción. No obstante, dicho resultado global, que se muestra debajo, sigue siendo un resultado más que aceptable. Pero además, según la estimación del porcentaje de acierto desglosado por clase, este sencillo árbol de decisión clasificaría correctamente casi la totalidad de las empresas «ma-

las» y, teniendo en cuenta que precisamente lo que nos interesa captar es la insolvencia, el modelo podría ser considerado como excelente en este sentido.

FIGURA 3.4. RESULTADOS *JACKKNIFE* ÁRBOL *MODELO 2*

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          53       77.9412 %
Incorrectly Classified Instances        15       22.0588 %
Total Number of Instances               68

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.941     0.382       0.711     0.941     0.81     mala
  0.618     0.059       0.913     0.618     0.737    buena

=== Confusion Matrix ===

  a  b   <-- classified as
 32  2 |  a = mala
 13 21 |  b = buena
```

Finalmente, emplearemos los datos del tercer año previo a la quiebra al objeto de predecir la crisis con tres años de antelación. A continuación se presenta el árbol de decisión seleccionado como *Modelo 3*, en el que se ha exigido que cada hoja cubra al menos 6 casos y se ha utilizado un *CF* del 25 % para la poda. Este árbol se muestra también en forma gráfica en el anexo 3.

FIGURA 3.4. RESULTADOS *JACKKNIFE* ÁRBOL *MODELO 2*

```
Test mode:    54-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

R17  <= 0.447703: buena (9.0/1.0)
R17  > 0.447703
|   R12  <= 0.534979: mala (7.0)
|   R12  > 0.534979
|   |   R6  <= 0.857601: mala (19.0/5.0)
|   |   R6  > 0.857601
|   |   |   R12  <= 2.262512: buena (11.0)
|   |   |   R12  > 2.262512: mala (8.0/3.0)

Number of Leaves  :     5

Size of the tree :      9
```

FIGURA 3.5. ÁRBOL DE DECISIÓN *MODELO 3*

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          53       77.9412 %
Incorrectly Classified Instances        15       22.0588 %
Total Number of Instances               68

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.941     0.382       0.711     0.941     0.81     mala
  0.618     0.059       0.913     0.618     0.737    buena

=== Confusion Matrix ===

  a  b   <-- classified as
 32  2 |  a = mala
 13 21 |  b = buena
```

Como se puede observar, cuando se trata de predecir el fracaso de las empresas de seguros con tres años de antelación las variables más discriminantes resultan ser *R17, R12* y de nuevo *R6.*

R17, que aparece en el nodo raíz, es el denominado «ratio de siniestralidad», considerado el indicador por excelencia de la eficiencia económica de un sistema de seguros, como se ha indicado en el capítulo anterior. Refleja el porcentaje de las primas que contablemente imputadas al ejercicio se consumen por los siniestros de la entidad:

$$R17 = \frac{\textit{Gastos técnicos seguro directo}}{\textit{Primas adquiridas seguro directo}}.$$

En el denominador figuran los ingresos adquiridos o periodificados, es decir, corregidos por la variación de provisiones, para, de este modo, correlacionar los ingresos que corresponden al ejercicio analizado con los siniestros del ejercicio cubiertos por los mismos.

Como refleja nuestro árbol de decisión, al señalar que las empresas cuya siniestralidad no supere el 45 % de las primas imputadas al ejercicio han de catalogarse como «buenas», será sintomático de eficiencia en la gestión técnica un valor bajo en este indicador. Sin embargo, cuando la siniestralidad sí supere ese 45 % sobre primas habrán de ser examinadas más variables para poder determinar la supervivencia o no de la empresa a tres años vista. En concreto, habría que estudiar en primer lugar el valor del ratio *R12.* Como ocurría con el *Modelo 2* a través del ratio *R14,* la solvencia vuel-

ve a manifestarse como factor clave a la hora de predecir el fracaso empresarial, salvo que en esta ocasión el denominador sólo incluye fondos propios.

En cualquier caso, cabría preguntarse por qué nuestro árbol señala que si *R12* es menor o igual de 0.53 —esto es, la siniestralidad no supera el 53 % de los fondos propios— la empresa será «mala», cuando en principio parecería solvente. La respuesta radicaría en la existencia de una escasa actividad en relación al volumen de fondos propios y la consiguiente caída de rentabilidad que, de no corregirse esta situación, finalmente desencadenaría la quiebra. Si, por contra, la siniestralidad sí superase el 53 % de los fondos propios, habría que examinar el ratio *R6,* el denominado «ratio de cobertura» que ya se manifestaba importante a la hora de predecir la crisis con dos y un año de antelación. En esta ocasión la empresa será «mala» cuando la cuantía de provisiones técnicas no supere el 86 % de la de ingresos por primas. En otro caso, vuelve a entrar en juego el ratio *R12,* de manera que si se verifican todas las condiciones que nos conducen a este último nodo concluiremos que la empresa será «buena» si la siniestralidad no supera el 226 % de los fondos propios y será «mala» en el caso contrario, ya que se trataría de una empresa descapitalizada en relación al volumen de siniestralidad.

En lo que respecta al poder predictivo de este árbol de decisión, en la clasificación de las 54 empresas utilizadas para la generación del mismo se obtiene un porcentaje global de acierto del 83 %, que se reduce al 81 % cuando es estimado mediante el método *jackknife*. Este resultado, que se muestra a continuación, nos conduce de nuevo a confiar en la bondad del modelo.

FIGURA 3.6. RESULTADOS *JACKKNIFE* ÁRBOL *MODELO 3*

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          44       81.4815 %
Incorrectly Classified Instances        10       18.5185 %
Total Number of Instances               54

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.926     0.296       0.758     0.926     0.833    mala
  0.704     0.074       0.905     0.704     0.792    buena

=== Confusion Matrix ===

  a  b   <-- classified as
 25  2 |  a = mala
  8 19 |  b = buena
```

Para finalizar este apartado, a modo de resumen se presenta debajo una tabla que recoge los porcentajes de acierto global y desagregado por clase de cada uno de los tres modelos, tanto los obtenidos sobre el propio conjunto de entrenamiento, como los estimados mediante la validación cruzada *jackknife*, así como el tamaño de los respectivos árboles y el conjunto de ratios intervinientes en los mismos.

Tabla 3.1. Resultados de los árboles de decisión

Modelo	Ratios	Tamaño del árbol	Clasificaciones correctas			
			Conjunto de entrenamiento		*Jackknife*	
			Empresas «buenas»	Empresas «malas»	Empresas «buenas»	Empresas «malas»
1	R3, R4, R5, R6	9	100 %	80,6 %	91,7 %	77,8 %
			Total: 90,3 %		Total: 84,7 %	
2	R14, R6, R3	7	61,8 %	94,1 %	61,8 %	94,1 %
			Total: 77,9 %		Total: 77,9 %	
3	R17, R12, R6	9	70,4 %	96,3 %	70,4 %	92,6 %
			Total: 83,3 %		Total: 81,5 %	

Como se puede apreciar, la precisión clasificatoria es menor en los años segundo y tercero previos a la quiebra que en el primero. Sin embargo, cabría preguntarse por qué dicha precisión es mayor en el tercer año que en el segundo, cuando la lógica normalmente indicaría que deberíamos esperar una disminución en la misma ante el aumento del horizonte temporal de la predicción. El caso es que este fenómeno se observa en ocasiones en otros estudios sobre predicción de insolvencias (Martínez de Lejarza Esparducer, 1999; Sanchis Arellano *et al*., 2003; Segovia Vargas, 2003) y no parece haber ninguna razón clara que lo justifique, con lo que no puede ser achacado más que a las peculiaridades de los datos, el reducido tamaño de la muestra y la eventual mala calidad de la información contable. No obstante, a pesar de lo anterior no existen grandes diferencias, sino que los resultados se muestran bastante estables en el tiempo, lo que podría significar que verdaderamente hemos encontrado en cada caso el subconjunto de ratios más relevantes para nuestro objetivo.

En cuanto a las variables incluidas en los modelos, podemos concluir que obviamente los más cercanos al momento de la crisis recogen algunas variables que reflejan problemas a corto plazo, mientras que cuanto más nos alejamos de dicho momento los ratios considerados reflejan problemas a más largo plazo. Tal es así para el caso del ratio de distancia a la quiebra contenido en el *Modelo 1*, ya que las empresas con anomalías observables en el valor de este ratio no sobrevivirían más de un ejercicio, o el ratio de liquidez contenido en los *Modelos 1* y *2*, que incluye partidas a corto plazo, y se muestra como más importante en el *Modelo 1* al ser considerado en el nodo raíz del árbol. En cambio, un factor que parece más relevante en el tercer año que en los anteriores, ya que sólo se refleja en el *Modelo 3*, es la eficiencia en la gestión técnica, mayor para las empresas «buenas», que deja entrever la mejor política de tarificación seguida en las mismas.

Otros factores, sin embargo, aparecen recogidos en varios modelos como diferenciadores entre empresas «buenas» y «malas», como el mayor volumen de provisiones técnicas en relación al nivel de actividad que poseen las empresas «buenas» indicado a través del ratio ***R6,*** que refleja una infradotación en las mismas en el caso de las empresas «malas», o el propio nivel de actividad, mayor para las «buenas», que se hace patente a través de los valores considerados para los ratios ***R5*** y ***R12***. Cabe señalar, asimismo, que por medio de las particiones realizadas en el tercer árbol en base a este último ratio queda puesta claramente de manifiesto la importancia de una cantidad equilibrada de fondos propios, de manera que cuantías tanto excesivas como escasas de los mismos en relación al volumen de siniestralidad serían sintomáticas de problemas a medio plazo, debido a su repercusión en la rentabilidad y en la solvencia dinámica de la empresa, respectivamente.

3.2.2. Reglas de clasificación

Aunque los árboles de decisión representan el conocimiento de manera muy sencilla, su inteligibilidad disminuye conforme aumenta su tamaño. Un conjunto de reglas de la forma «si» (condiciones) - «entonces» (decisión) es un mecanismo de representación del conocimiento más inteligible

que los árboles de decisión, puesto que cuando el problema es complejo el árbol generado es tan grande que ni siquiera tras su poda resulta sencillo comprender el modelo de clasificación completo, y por ello las reglas de clasificación son una alternativa muy popular a los árboles de decisión.

El «antecedente» o conjunto de condiciones de una regla, al igual que los nodos internos de un árbol de decisión, contiene una serie de preguntas, mientras que el «consecuente» o conclusión indica la clase de las instancias cubiertas por esa regla, o quizá una distribución de probabilidad sobre las clases.

Si bien un árbol de decisión también puede ser representado como un conjunto de reglas de manera trivial, puesto que de cada camino desde la raíz del árbol hasta una hoja se deriva una regla, este modo de obtener una regla por cada hoja del árbol, sin embargo, no es más que otra manera de describir el mismo y, por tanto, tales reglas seguirán siendo mutuamente excluyentes. Lo interesante sería poder simplificar estas reglas generadas, ya que pueden contener condiciones irrelevantes en su antecedente que podrían eliminarse sin que disminuyera la precisión del clasificador.

Como se ha explicado con anterioridad en este trabajo, el algoritmo C4.5 implementa un método para llevar a cabo la simplificación de las reglas derivadas de un árbol de decisión no pospodado así como reducir el número de éstas y, dado que, de este modo, se perderán la exclusividad y exhaustividad características de las condiciones que integran un árbol de decisión, también se implementan mecanismos para resolver los conflictos que se presenten cuando existan casos cubiertos por varias reglas correspondientes a clases distintas y para contemplar la situación de que existan casos no cubiertos por ninguna de las reglas.

Procedamos entonces a derivar sendos conjuntos de reglas a partir de los ratios de las empresas de nuestra muestra para los tres años previos a la quiebra, tratando siempre de evitar la generación de grupos de reglas muy complejos que clasifiquen perfectamente el conjunto de entrenamiento pero, sin embargo, manifiesten un mal comportamiento en la validación cruzada y, por tanto, una escasa capacidad de generalización.

Tomando los datos del primer año anterior a la quiebra y tras varias combinaciones de los parámetros de aprendizaje, siendo éstos el

parámetro *m* que controla el número mínimo de instancias que ha de cubrir cada hoja del árbol de partida y el parámetro *CF* fijado para el proceso de simplificación de reglas (es decir, el factor de confianza usado para podar reglas en vez de árboles), hemos seleccionado como *Modelo 1* el conjunto de reglas que se muestra en la siguiente figura (figura 3.7), el cual se ha obtenido con los valores 3 y 1 % para *m* y *CF,* respectivamente.

FIGURA 3.7. REGLAS DE CLASIFICACIÓN *MODELO 1*

```
Rule 1:
        R3 <= 1.40189
        R4 <= 1.46373
        ->  class mala  [78.5%]

Rule 4:
        R6 <= 0.0493137
        ->  class mala  [59.9%]

Rule 5:
        R3 > 1.40189
        R6 > 0.0493137
        ->  class buena  [63.5%]

Rule 2:
        R4 > 1.46373
        R4 <= 1.86373
        ->  class buena  [59.9%]

Default class: mala


Evaluation on training data (72 items):

Rule  Size  Error  Used  Wrong            Advantage
----  ----  -----  ----  -----            ---------
   1     2  21.5%    19      0 (0.0%)        0 (0|0)    mala
   4     1  40.1%     7      0 (0.0%)        0 (0|0)    mala
   5     2  36.5%    40      7 (17.5%)      20 (27|7)   buena
   2     2  40.1%     3      0 (0.0%)        3 (3|0)    buena

Tested 72, errors 7 (9.7%)   <<

          (a)   (b)      <-classified as
          ----  ----
            29     7      (a): class mala
                  36      (b): class buena
```

Como se puede apreciar en la figura, el modelo consta de cuatro reglas (dos para cada una de las clases) y la clase por defecto (en este caso «mala»). El número de cada regla es arbitrario, se deriva del orden de

las hojas en el árbol original, y tan sólo representa un modo de identificar las reglas. El antecedente de cada regla se integra por una serie de condiciones relativas a los valores de ciertos atributos que han de ser todas satisfechas para que la regla sea aplicable, y el consecuente indica la decisión a tomar. Al lado, entre corchetes, se refleja la precisión estimada de la regla (que se obtendrá como $1-p_e$, siendo p_e la estimación de la probabilidad de error que se realiza para llevar a cabo el proceso de simplificación comentado en el primer capítulo del presente trabajo). Así, en la primera regla (*Rule* 1), se prevé que la clasificación será correcta para el 78.5 % de nuevos casos que satisfagan las condiciones de la regla.

Debajo del conjunto de reglas se presenta la evaluación de cada una de ellas sobre los casos de entrenamiento. Por ejemplo, para la tercera (*Rule* 5), se indica que el antecedente se compone de dos condiciones, la tasa de error prevista p_e es del 36.5 %, se usó 40 veces en la clasificación de los casos de entrenamiento y 7 fueron errores. La columna titulada «*Advantage*» muestra qué ocurriría si la regla se omitiese del conjunto. Cada entrada de la forma $a\ (b|c)$ indica que, si la regla fuera omitida, b casos clasificados ahora correctamente por esta regla serían clasificados incorrectamente, y c casos ahora mal clasificados por esta regla serían clasificados correctamente por las reglas siguientes y la clase por defecto; el beneficio neto de no omitir la regla es entonces $a = b - c$.

Por último, se muestra el número total de errores (incluyendo las clasificaciones erróneas de la clase por defecto), así como una matriz de confusión que señala el tipo de errores cometidos.

Para evaluar la capacidad predictiva del modelo, dado que el elevado porcentaje de clasificaciones correctas (90.3 %) obtenido sobre el propio conjunto de datos usado para la construcción de las reglas no representa una medida adecuada de la validez del modelo de cara a la clasificación de nuevas empresas, llevamos a cabo el procedimiento de validación *jackknife*, obteniendo los siguientes resultados:

FIGURA 3.8. RESULTADOS *JACKKNIFE* REGLAS DE CLASIFICACIÓN - *MODELO 1*

```
train:  Tested 71.0, errors 6.8 (9.6%)
test:   Tested 1.0, errors 0.2 (15.3%)
```

Como se puede observar, la media de los errores de los 72 modelos parciales construidos con 71 casos es del 9.6 % (lo que representa un 90.4 % de aciertos en la clasificación de los casos de entrenamiento) y el error medio al evaluar dichos modelos con los casos de test es del 15.3 % (84.7 % de aciertos). De este modo, el conjunto de reglas que representa el *Modelo 1*, además de ser un clasificador más simple que el árbol de decisión obtenido para el primer año anterior a la quiebra, se puede considerar tan fiable como el mismo a la hora de utilizarlo para clasificar otras empresas, en cuyo caso se aplicaría primero cualquiera de las reglas representativas de la clase ordenada en primer lugar («mala») y si ninguna de ellas fuera aplicable se pasaría a las reglas correspondientes a la clase ordenada en segundo lugar («buena»); si éstas tampoco fuesen aplicables, a la empresa se le asignaría la clase por defecto («mala»).

El antecedente de la primera de las reglas (*Rule* 1) refleja dos condiciones que se han de cumplir para que las empresas sean asignadas a la clase «mala» designada en el consecuente. Dichas condiciones se refieren al ratio de liquidez ***R3*** y al ratio de garantía o distancia a la quiebra ***R4.*** Esta regla coincide con una de las reglas derivadas del árbol de decisión obtenido para el primer año anterior a la quiebra, e indica que si el activo circulante de la empresa no supera el 140 % del pasivo circulante y el activo real es menor o igual que el 146 % del pasivo exigible, la empresa será «mala».

La segunda de las reglas (*Rule* 4) se integra por una única condición relativa al ratio ***R6,*** y señala que las empresas con un nivel de provisiones técnicas inferior al 5 % de los ingresos por primas serán «malas». Esta condición también aparecía en el árbol, aunque no aislada como ahora sino anidada con otra condición relativa al ratio de liquidez que, como podemos comprobar en este momento, puede ser eliminada sin pérdida de precisión clasificatoria, puesto que las empresas con un valor tan bajo en el ratio ***R6*** han de considerarse «malas» sin necesidad de observar el valor de ningún otro ratio.

La tercera regla (*Rule* 5) es otra de las existentes en el árbol, y refleja que para que la empresa sea «buena» no basta con que el ratio ***R6*** supere el 4,9 % sino que además su ratio de liquidez debe ser mayor del 140 %.

Para la cuarta regla (*Rule* 2) no existe una homóloga en el árbol. En éste la condición *R4* > 1.46 se anidaba con el ratio de liquidez y con el ratio *R5,* que no aparece en el conjunto de reglas, y en su lugar el nivel de actividad es representado a través de la segunda de las condiciones de esta cuarta regla, ya que teniendo en cuenta que el principal componente del pasivo en las empresas de seguros está constituido por provisiones técnicas, valores muy elevados en el ratio *R4* se deberían a un nivel bajo de dichas provisiones técnicas que puede ser consecuencia tanto de una constitución incorrecta de las mismas, como de un escaso volumen de actividad, que, en cualquier caso, no recomendarían clasificar la empresa como «buena». Por ello la regla establece un límite para el valor del ratio *R4*, de manera que si éste superase el 186 % ya no podría aplicarse esta regla representativa de la clase «buena».

FIGURA 3.9. REGLAS DE CLASIFICACIÓN *MODELO 2*

```
Rule 5:
        R3 <= 1.61196
        R14 > 0.399419
        ->  class mala  [51.5%]

Rule 6:
        R3 > 1.61196
        R14 <= 0.633184
        ->  class buena  [57.8%]

Rule 4:
        R6 > 0.906477
        R14 <= 0.399419
        ->  class buena  [56.4%]

Default class: mala


Evaluation on training data (68 items):

Rule  Size  Error  Used  Wrong            Advantage
----  ----  -----  ----  -----            ---------
   5     2  48.5%    20      4 (20.0%)       0 (0|0)    mala
   6     2  42.2%    24      4 (16.7%)      12 (15|3)   buena
   4     2  43.6%     7      0 (0.0%)        7 (7|0)    buena

Tested 68, errors 11 (16.2%)   <<


         (a)   (b)      <-classified as
        ----  ----
          30     4      (a): class mala
           7    27      (b): class buena
```

Finalmente, según establece la regla por defecto, habrá que considerar «malas» aquellas empresas para las cuales no sea aplicable ninguna de las reglas anteriores.

Utilizando ahora los datos del segundo año previo a la quiebra, hemos seleccionado como *Modelo* 2 el conjunto de reglas mostrado en la siguiente figura, que se ha obtenido con un valor de 2 para el parámetro *m* y 1 % para el parámetro *CF*.

Como se aprecia en la figura 3.9, el porcentaje de acierto en la clasificación de las 68 empresas que constituyen el conjunto de entrenamiento es del 83.8 %. En cuanto a la validación cruzada del modelo, a continuación se muestran los resultados de la misma, donde se puede observar que el error medio al evaluar los 68 modelos parciales con los casos de test es del 22.1 %, lo que supone un porcentaje de acierto estimado en la clasificación de nuevas empresas del 77.9 %.

FIGURA 3.10. RESULTADOS *JACKKNIFE* REGLAS DE CLASIFICACIÓN - *MODELO 2*

```
train:  Tested 67.0, errors 10.8 (16.1%)
test:   Tested 1.0, errors 0.2 (22.1%)
```

Como muestran los resultados obtenidos en la validación cruzada, el modelo representado por medio de este sencillo conjunto de reglas es tan preciso como el árbol de decisión derivado para el segundo año anterior a la quiebra, pero además, aunque las variables intervinientes en ambos modelos son las mismas, la representación por medio de reglas resulta ser más sencilla pues consta de un menor número de condiciones que las habidas en el árbol.

Por otro lado, sólo una de las reglas del conjunto se reflejaba en el árbol, la «*Rule* 4» que señala que si la cuantía de provisiones técnicas es mayor que el 91 % de la de ingresos por primas y además la siniestralidad no supera el 40 % de la suma de fondos propios y provisiones técnicas, la empresa será «buena». Sin embargo, para la «*Rule* 5» (si el activo circulante es menor o igual que el 161 % del pasivo circulante y la siniestralidad es superior al 40 % de la suma de fondos propios y provisiones técnicas la empresa es «mala») y la «*Rule* 6» (cuando el activo circulante sea mayor que el 161 % del pasivo circulante y la siniestralidad no supere el 63 % de la suma de fondos propios y provi-

siones técnicas la empresa será «buena») no existen reglas homólogas en el árbol.

Recordemos que para clasificar una nueva empresa se comenzará comprobando las condiciones del conjunto de reglas representativo de la clase situada en primer lugar («mala» en nuestro modelo) y si ninguna de dichas reglas fuese aplicable se probaría con el grupo de reglas de la clase ordenada en segundo lugar, y así sucesivamente, aunque en nuestro caso, como el problema sólo consta de dos clases, sólo habrá dos grupos de reglas. Finalmente, se añade una regla o clase por defecto para cuando ninguna de las reglas anteriores sea aplicable, y que será «mala» según el *Modelo 2*.

FIGURA 3.11. REGLAS DE CLASIFICACIÓN *MODELO 3*

```
Rule 4:
        R6 > 0.857601
        R12 > 0.534979
        R12 <= 2.26251
        ->  class buena  [76.2%]

Rule 1:
        R17 <= 0.447703
        ->  class buena  [56.0%]

Rule 2:
        R12 <= 0.534979
        R17 > 0.447703
        ->  class mala  [65.2%]

Rule 3:
        R6 <= 0.857601
        R17 > 0.447703
        ->  class mala  [57.6%]

Default class: mala

Evaluation on training data (54 items):

Rule  Size  Error  Used  Wrong            Advantage
----  ----  -----  ----  -----            ---------
   4     3  23.8%    11      0 (0.0%)       11 (11|0)   buena
   1     1  44.0%     9      1 (11.1%)       7 (8|1)    buena
   2     2  34.8%     7      0 (0.0%)        0 (0|0)    mala
   3     2  42.4%    19      5 (26.3%)       0 (0|0)    mala

Tested 54, errors 9 (16.7%)   <<

         (a)   (b)      <-classified as
        ----  ----
          26     1      (a): class mala
           8    19      (b): class buena
```

Tratando ahora de predecir la crisis con tres años de antelación hemos seleccionado como *Modelo 3* el conjunto de reglas que se expone en la siguiente figura (figura 3.11), el cual se ha obtenido con los valores 6 y 5 % para los parámetros *m* y *CF*, respectivamente.

Como se puede observar en la figura 3.11, el porcentaje global de acierto sobre el conjunto de entrenamiento es del 83.3 %, y en cuanto a los resultados de la validación cruzada *jackknife*, que se muestran a continuación (figura 3.12), se obtiene un error medio al evaluar los 54 modelos parciales con los casos de test del 18.5 %, lo que supone un porcentaje de acierto estimado del 81.5 %.

FIGURA 3.12. RESULTADOS *JACKKNIFE* REGLAS DE CLASIFICACIÓN - *MODELO 3*

```
train:  Tested 53.0, errors 8.8 (16.6%)
test:   Tested 1.0, errors 0.2 (18.5%)
```

Nuestro conjunto de reglas resulta ser de nuevo un clasificador tan fiable como el árbol de decisión obtenido para el tercer año previo a la quiebra, pero además más sencillo. Así, sólo dos de las reglas coinciden con sendas reglas de las cinco existentes en el árbol, siendo éstas las «*Rule* 1» y «*Rule* 2» que señalan, respectivamente, que una empresa será «buena» cuando la siniestralidad no supere el 45 % de las primas imputadas al ejercicio, y será «mala» si además de ser el ratio de siniestralidad mayor que el 45 % dicha siniestralidad no supera el 53 % de la cuantía de fondos propios.

Sin embargo, la «*Rule* 4» es más sencilla que la regla del árbol donde aparecen también estas tres condiciones, ya que en aquél dichas tres condiciones se anidan con otra relativa al ratio de siniestralidad, que ha resultado ser innecesaria al ser eliminada ahora sin pérdida de eficacia. Según esta regla «*Rule* 4» cuando la cuantía de provisiones técnicas sea mayor que el 86 % de la de ingresos por primas y además el ratio *R12* sea mayor que 0.53 pero no superior a 2.26 la empresa será «buena», quedando patente de nuevo a través del intervalo considerado para el ratio *R12* la importancia de una cuantía equilibrada de fondos propios en relación al volumen de siniestralidad.

Por su parte, la «*Rule* 3» también es más sencilla que la regla del árbol donde aparecen reflejadas estas dos condiciones relativas al ratio

de cobertura *R6* y al ratio de siniestralidad *R17*, que se anidaban en el árbol con el ratio *R12*. De este modo, cuando la cuantía de provisiones técnicas sea menor o igual que el 86 % de la de ingresos por primas y, asimismo, la siniestralidad supere el 45 % de las primas imputadas al ejercicio, la empresa será «mala», considerándose también «malas» aquellas empresas para las cuales no sea aplicable ninguna de las reglas anteriores.

Finalmente, en la siguiente tabla se resumen los resultados alcanzados por cada uno de los tres modelos, recogiéndose los porcentajes de acierto global y desagregado por clase obtenidos sobre los conjuntos de entrenamiento y los porcentajes de acierto global estimados mediante la validación cruzada *jackknife* (en este caso no se presenta la desagregación por clase debido a que el programa no la suministra), así como el número de reglas de que consta cada modelo (incluyendo la clase por defecto) y el conjunto de ratios intervinientes en los mismos.

Tabla 3.2. Resultados de las reglas de clasificación

Modelo	Ratios	Número de reglas	Clasificaciones correctas		
			Conjunto de entrenamiento		*Jackknife*
			Empresas «buenas»	Empresas «malas»	
1	R3, R4, R6	5	100 %	80,6 %	
			Total: 90,3 %		Total: 84,7 %
2	R3, R14, R6	4	79,4 %	88,2 %	
			Total: 83,8 %		Total: 77,9 %
3	R6, R12, R17	5	70,4 %	96,3 %	
			Total: 83,3 %		Total: 81,5 %

3.2.3. Comparación con Regresión Logística

La Regresión Logística surge como una extensión de la Regresión Lineal ordinaria basada en el método de los mínimos cuadrados para superar las limitaciones de esta técnica cuando es utilizada con variables dependientes categóricas (Peña, 2002). Adicionalmente, presenta la ventaja frente al Análisis Discriminante de no requerir el cumplimiento de las estrictas hipótesis acerca de la distribución de las variables que justifican (al menos en teoría) la aplicación de esta última herramienta.

La Regresión Logística consiste en realizar una estimación por máxima verosimilitud de los parámetros de una función lineal de las variables explicativas. El modelo planteado tendrá la forma:

$$log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon,$$

donde ε es el término de error y p la probabilidad de éxito en una variable aleatoria binaria que sigue una distribución de Bernoulli. Los valores que toma esta variable indican la clase a la que pertenece cada observación. Dada una nueva observación caracterizada por unos valores concretos de $x_1, x_2, \ldots, x_p$, el modelo nos da la probabilidad estimada de que esa observación pertenezca a una u otra clase. En un problema de clasificación la observación será asignada a la clase más probable de acuerdo con los valores proporcionados por el anterior modelo.

Un problema que plantea esta técnica es su incapacidad para aceptar valores faltantes, es decir, la matriz de datos que se le suministre debe tener todos sus valores conocidos. Sin embargo, en nuestras matrices de datos en bruto existe un pequeño número de *missing values* (6 para el primer año previo a la quiebra y 3 para cada uno de los otros dos años) consecuencia de que, como se ha indicado anteriormente, se ha optado por considerar como valores faltantes los de los ratios cuyo denominador resulta ser cero (y, por tanto, el valor del ratio infinito).

Existe otro problema relacionado con el anterior y es que, como revela una sencilla inspección ocular, hay una serie de valores dentro de las matrices de datos que son extraordinariamente atípicos (por ejemplo, en el ratio *R5* para el año 1 existe un valor que es 50 veces más grande que el inmediatamente inferior). Aunque son pocos (aproximadamente un 1 % del total), tales valores distorsionan enormemente los resultados obtenidos con la Regresión Logística, por lo que se ha optado por eliminarlos convirtiéndolos en *missing values*. Para determinar con una cierta objetividad qué valores son lo suficientemente extremos como para justificar esta eliminación se ha procedido a estandarizar los datos, restándole a cada uno de ellos la media de la correspondiente variable y dividiendo el resultado también por la correspondiente desviación típica. Aquellos valores que así es-

tandarizados resultan tener un valor absoluto superior a 100 son considerados extremos y convertidos en *missing values*. La media y la desviación típica que se utilizan en este proceso de estandarización no son los valores habituales, sino medidas robustas de posición y dispersión, ya que así lo aconseja la presencia de valores tan atípicos como los indicados. Así, en lugar de la media, se toma una media recortada *(trimmed mean)* en la que se eliminan el 10 % superior e inferior de los valores. En lugar de la desviación típica se toma la desviación absoluta media, que es el valor absoluto de la mediana de la variable centrada con respecto a su mediana y multiplicado todo ello por un factor que la convierte en un estimador insesgado de la desviación típica cuando esta magnitud se extrae de una variable aleatoria con distribución normal (R Development Core Team, 2005). Tanto la media recortada como la desviación absoluta media son estimadores de notable robustez frente a la presencia de datos atípicos y es por ello por lo que han sido utilizadas.

Con este procedimiento se convierte, como se ha indicado, aproximadamente un 1 % de valores en *missing values*. Posteriormente, dichos valores serán imputados. En lugar de imputar los *missing values* se podría haber optado por desechar aquellos casos para los cuales alguno de los 25 ratios sea un valor faltante. Sin embargo, el tamaño de la muestra es pequeño y esto la reduciría aún más y supondría descartar información útil simplemente porque alguno de los valores de un caso sea *missing* (lo cual sólo parece razonable con tamaños de muestra grandes), de modo que finalmente se ha optado por el tratamiento descrito.

Si bien el procedimiento de imputación más habitual consiste en utilizar la media o la mediana del resto de valores de la variable en cuestión, hemos elegido una alternativa un tanto más laboriosa y realista que se describe en Troyanskaya *et al*. (2001). En este artículo se comparan varios métodos de imputación comprobándose que el que proporciona mejores resultados es el denominado *KNNimpute*, que ha sido por tanto el que hemos puesto en práctica.

El método *KNNimpute* consiste en seleccionar para cada caso con algún valor faltante los k casos más cercanos a él con todos sus valores

completos (serán los k vecinos más próximos al caso con el valor faltante). La proximidad se medirá con la distancia euclídea, aunque son posibles otras alternativas que serían más adecuadas si los datos estuviesen expresados en términos absolutos (no como ratios), ya que en este caso el efecto de las unidades de medida podría alterar enormemente los resultados. Una vez determinados los k vecinos más próximos al caso con el valor faltante, este valor se imputará tomando la media ponderada de acuerdo con la distancia de los valores correspondientes de esos k vecinos.

Para determinar el número de vecinos más adecuado, es decir, el valor óptimo de k, se ha seguido un procedimiento expuesto también en Troyanskaya *et al*. (2001) y que consiste en realizar una serie de simulaciones que permiten averiguar ese valor de k. Para ello se parte de una matriz de datos completa (eliminando en la matriz original todos los casos con algún valor faltante) y sobre ella se elimina aleatoriamente un pequeño porcentaje de valores para convertirlos en *missing values*. Estos valores eliminados son a continuación imputados con diferentes valores de k comparando las matrices así imputadas con la matriz completa de partida. La comparación se realiza tomando la raíz cuadrada de la media del cuadrado de la diferencia entre los elementos de las matrices comparadas (la original con la imputada) y dividiendo esta cantidad por el valor medio de la matriz completa (se obtendrá entonces un error cuadrático medio normalizado). El valor de k para el cual esta cantidad sea mínima en promedio sobre un número suficientemente grande de simulaciones como para obtener unos resultados razonablemente estables será el que se utilice finalmente para realizar la imputación. Para automatizar este proceso se ha escrito una pequeña función en R cuyo código se incluye en el anexo 4. Los resultados obtenidos para los tres años se recogen en el anexo 5, donde se puede apreciar cómo el valor adecuado de k es 8 en todos los casos.

Otro problema es el derivado de la necesidad de determinar cuáles serán las variables explicativas más adecuadas de entre el conjunto de 25 ratios que se incluirán en los modelos de Regresión Logística. Con esta selección previa de las variables se consigue eliminar problemas de

colinealidad que harían inestables los resultados, obtener modelos más sencillos y fáciles de interpretar y reducir el sobreajuste. Un enfoque habitual es el constituido por los procedimientos de tipo *stepwise* que utilizan contrastes de significatividad basados en las distribuciones de la t de Student y la F de Snedecor. Sin embargo, tales procedimientos son intrínsecamente inestables y dependen en buena medida del cumplimiento de hipótesis bastante estrictas acerca de la distribución de las variables consideradas. Ello nos ha llevado a optar por el denominado *Bayesian Information Criterion* (BIC), que utiliza ideas procedentes de la Teoría de la Información para seleccionar aquel modelo que minimice la expresión $-2 \log [L (\hat{\theta})] + p \log n$, en la que n es el número de observaciones, p, el número de variables y $\hat{\theta}$, el estimador máximo verosímil de los parámetros del modelo. Este criterio tiende a seleccionar modelos muy aceptables en el caso de pequeños tamaños muestrales y cuenta con un notable respaldo teórico (Peña, 2002).

En la siguiente tabla se presentan los ratios seleccionados de acuerdo con el mencionado criterio para construir cada modelo, así como los resultados obtenidos en la clasificación.

Tabla 3.3. Resultados Regresión Logística

Modelo	Ratios	Clasificaciones correctas			
		Conjunto de entrenamiento		*Jackknife*	
		Empresas «buenas»	Empresas «malas»	Empresas «buenas»	Empresas «malas»
1	R5, R6, R7, R8, R12, R23, R24	83,3 %	77,8 %	80,6 %	66,7 %
		Total: 80,6 %		Total: 73,6 %	
2	R5, R9, R10, R11, R12, R23	82,4 %	67,6 %	76,5 %	61,8 %
		Total: 75 %		Total: 69,1 %	
3	R4, R6, R10, R11, R17 R19, R20, R22, R24	70,4 %	77,8 %	66,7 %	66,7 %
		Total: 74,1 %		Total: 66,7 %	

La información relativa a los coeficientes y la significatividad de cada uno de los modelos construidos aparece recogida a continuación:

Año 1

```
Call:
glm(formula = clase ~ R5 + R6 + R7 + R8 + R12 + R23 + R24,
        family = binomial(link = logit), data = anno1sss)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.89374  -0.70351   0.01141   0.73468   2.30653

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6561     0.6483   2.555  0.01063 *
R5           -1.1827     0.5213  -2.269  0.02329 *
R6           -2.6247     1.4438  -1.818  0.06907 .
R7            2.1685     1.2735   1.703  0.08859 .
R8            2.0072     0.6404   3.134  0.00172 **
R12          -1.6063     0.5229  -3.072  0.00213 **
R23          -8.3867     3.4650  -2.420  0.01551 *
R24          19.5242     6.5938   2.961  0.00307 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99.813  on 71  degrees of freedom
Residual deviance: 64.634  on 64  degrees of freedom
AIC: 80.634

Number of Fisher Scoring iterations: 7
```

Año 2

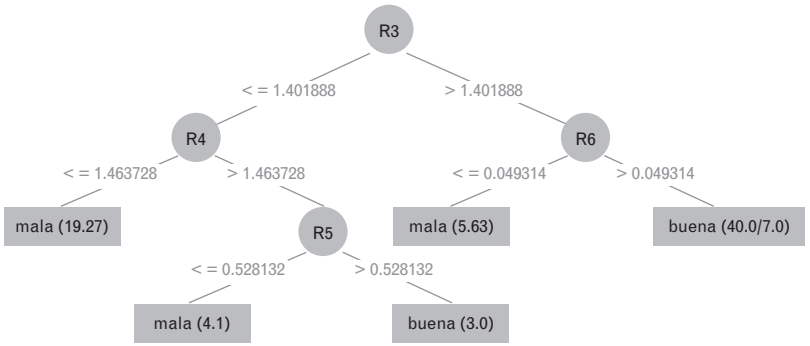
```
Call:
glm(formula = clase ~ R5 + R9 + R10 + R11 + R12 + R23,
    family = binomial(link = logit), data = anno2sss)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.90351  -0.66640   0.01591   0.80330   2.41160

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1903     0.4899   2.430  0.01510 *
R5           -1.7040     0.6403  -2.661  0.00778 **
R9            2.4675     0.8538   2.890  0.00385 **
R10           3.5726     1.8450   1.936  0.05282 .
R11          -3.1401     1.7157  -1.830  0.06722 .
R12          -1.6030     0.6981  -2.296  0.02167 *
R23           4.0205     1.2338   3.259  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 94.268  on 67  degrees of freedom
Residual deviance: 63.331  on 61  degrees of freedom
AIC: 77.331

Number of Fisher Scoring iterations: 6
```

Año 3

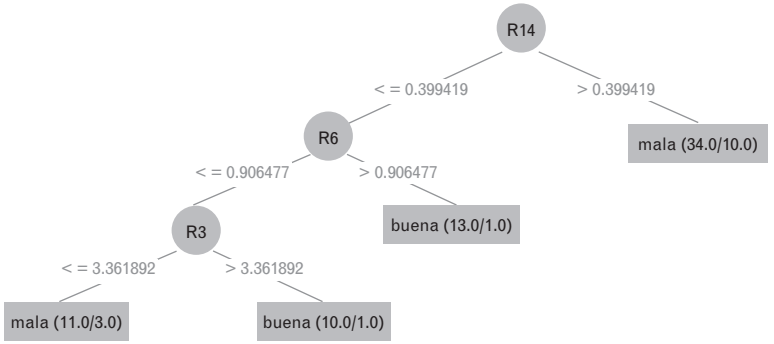
```
Call:
glm(formula = clase ~ R4 + R6 + R10 + R11 + R17 + R19 + R20 +
    R22 + R24, family = binomial(link = logit), data = anno3sss)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.75248  -0.71655   0.00924   0.66166   1.90983

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.3157     2.7793  -1.553  0.12047
R4            0.9343     0.3653   2.557  0.01055 *
R6            3.0526     1.4096   2.166  0.03034 *
R10           2.0657     0.7634   2.706  0.00681 **
R11          -1.9865     0.7832  -2.537  0.01120 *
R17         -17.0743     6.7015  -2.548  0.01084 *
R19         -13.6835     5.2669  -2.598  0.00938 **
R20          16.1500     6.7687   2.386  0.01703 *
R22         -14.2711     6.9952  -2.040  0.04134 *
R24          14.5405     7.9646   1.826  0.06790 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.860  on 53  degrees of freedom
Residual deviance: 48.588  on 44  degrees of freedom
AIC: 68.588

Number of Fisher Scoring iterations: 5
```

3.2.4. Anexos

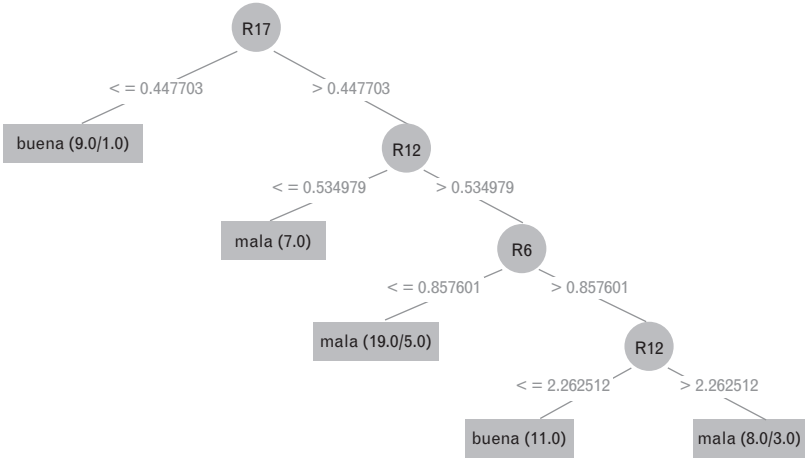
Anexo 1. Árbol de decisión *modelo 1*



Anexo 2. Árbol de decisión *modelo 2*



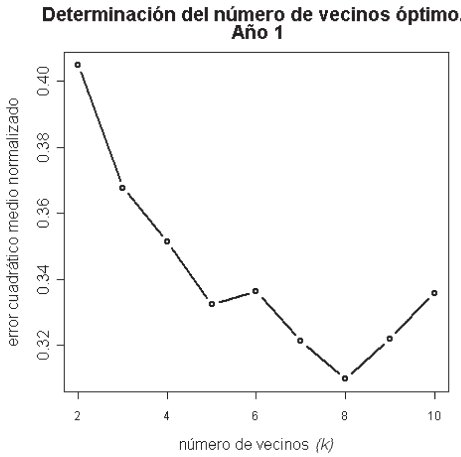
Anexo 3. Árbol de decisión *modelo 3*

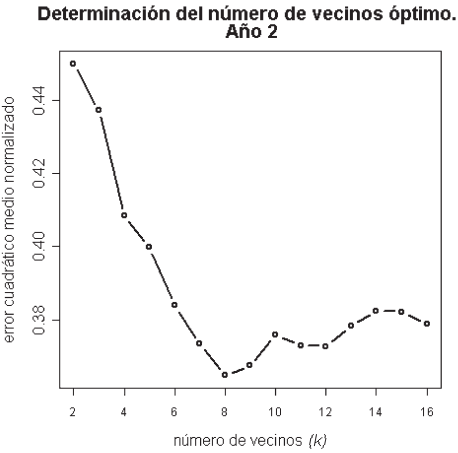


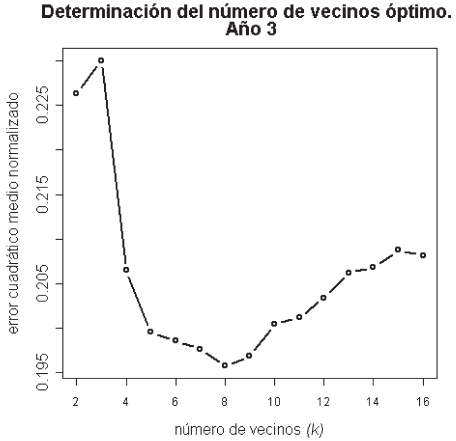
Anexo 4. Función usada para la determinación del *K* óptimo

```
sacark5<-function(mmm,k,percen=0.01,iter=50) {

      # mmm será la matriz (puede tener missing porque se filtra)
      # k será un vector con los valores de k que se quieren probar
      # percen será el porcentaje de casos de mmm que se hacen missing:por defecto el 1 %
      # iter será el número de veces que se repite el proceso para cada  k: por defecto se toma 50
      # la función devuelve una lista con dos componentes:
      # el vector k y el error cuadrático medio que se obtiene al imputar para cada valor de K

 mm <- na.omit(mmm);
      lk<-length(k);
      rr<-matrix(0,lk,1);
      ss<-matrix(0,lk,1);
      mediat<-mean(mm);
      dimen<-dim(mm);
      nele<-prod(dimen);

      for (i in 1:iter)
            {
                  dd<-sample(nele,round(percen*nele));
                  dd<-sort(dd);
                  a1cna<-mm;
                  a1cna[dd]<-NA;
                  a1cna<-matrix(a1cna,dimen[1],dimen[2]);
                  # a1cna será una matriz con missing aleatorios

                  for (j in 1:lk)
                        {
                        ma<-a1cna;
                        ma<-knn(ma,k[j])$data;
                        aa<-(mm-ma);
                        ss[j]<-(sqrt(mean(aa^2)))/mediat;
                        };
                  rr<-rr+ss;
            };
      rr<-rr/iter;
      # rr almacena el error cuadrático
      list(kk=k,rms=rr)
}
```

Anexo 5. Valor de *K* óptimo




Determinación del número de vecinos óptimo.
Año 2
error cuadrático medio normalizado
0.38
0.40
0.42
0.44
2
4
6
8
10
12
14
16
número de vecinos (k)


Determinación del número de vecinos óptimo.
Año 3
error cuadrático medio normalizado
0.195
0.205
0.215
0.225
2
4
6
8
10
12
14
16
número de vecinos (k)

Capítulo 4:
Conclusiones

Concluida la parte empírica no cabe ya sino que expongamos a modo de conclusión una serie de reflexiones que se han ido suscitando a lo largo del desarrollo de este trabajo.

Lo que con él hemos pretendido es mostrar la adecuación de un paradigma procedente del área de la Inteligencia Artificial conocida como Aprendizaje Automático, el algoritmo de inducción de árboles de decisión y reglas de clasificación C4.5, para la valoración de la solvencia de las entidades aseguradoras. Para ello acudimos al terreno empírico, utilizando un conjunto de ratios financieros calculados a partir de los estados contables de una muestra de sociedades anónimas españolas de seguros no vida. Al objeto de tener una referencia que pueda ser utilizada como término de comparación, aplicamos también a nuestra muestra Regresión Logística por ser ésta una técnica estadística estándar.

A la luz de los resultados obtenidos queda puesta claramente de manifiesto la superioridad del algoritmo C4.5 a la hora de predecir el fracaso empresarial en sociedades españolas de seguros no vida. Este método utiliza de forma más eficiente la información disponible que la técnica estadística, lo cual conduce a una tasa de clasificación correcta más alta. Probablemente la estructura del espacio de datos sea demasiado compleja para poder lograr una buena separación de forma lineal, y el modo más sofisticado en que el algoritmo de aprendizaje automático lleva a cabo la separación entre las clases se adapte mejor a la estructura inherente a los datos.

Por otro lado, además de en porcentaje de acierto, el algoritmo C4.5 supera a la técnica estadística también en otros aspectos: se aplica con

mayor facilidad, proporciona modelos de interpretación más sencilla y es robusto ante el «ruido» introducido por valores faltantes y *outliers,* lo que hace especialmente atractiva su utilización para el caso de la información contable, que suele presentar datos interrelacionados, incompletos, adulterados o erróneos.

Respecto al objetivo que nos planteábamos con la realización de este trabajo consistente en demostrar la adecuación de C4.5 al problema concreto de la predicción del fracaso empresarial en las empresas españolas de seguros no vida, pensamos que ha sido satisfactoriamente alcanzado.

Ahora bien, es importante señalar que la muestra de empresas utilizada para llevar a cabo nuestro estudio empírico, además de ser relativamente pequeña, abarca datos del período comprendido entre 1983 y 1993, por lo que, teniendo en cuenta las notables transformaciones que desde entonces ha sufrido el sector en su regulación, estructura y funcionamiento, extrapolar los resultados a día de hoy no sería aconsejable en modo alguno. Por ello, en la medida en que se disponga de datos más actuales sería conveniente elaborar nuevos modelos incorporando además otras variables predictoras imposibles de extraer a partir de la información contable de entonces, como por ejemplo, aquéllas relativas al margen de solvencia u otras que se puedan obtener de los modelos de cuentas anuales obligatorios según las normas recogidas en el Plan de Contabilidad de las entidades aseguradoras actualmente vigente.

También conviene tener presente que el fracaso empresarial es el resultado de un proceso en el que interactúan muchos más factores además de las variables estrictamente financieras consideradas en nuestro trabajo, tanto de carácter interno como externo a la propia empresa. Aunque a nosotros no nos fue posible, debido a que sólo contábamos con los balances y las cuentas de pérdidas y ganancias de las empresas de la muestra, si se dispusiese de ella sería interesante incorporar a los modelos información de tipo cualitativo que el algoritmo es capaz de manejar y que probablemente mejoraría la capacidad de predicción.

Por otro lado, en la obtención de los modelos hemos perseguido en todo momento la minimización del porcentaje de clasificación errónea. Sin embargo, la autoridad supervisora podría estar más interesada en

minimizar los costes que los errores implican. Obviamente, será más importante el coste de clasificar como sana una empresa que en realidad será quebrada que el de catalogar como fracasada una empresa sana, ya que lo primero supondría «dejar pasar» la oportunidad de anticiparse y salvar a una empresa de la quiebra y evitar los efectos perniciosos de la misma. En la medida en que se disponga de estimaciones razonables de estos costes podrían ser incorporados al análisis, puesto que el algoritmo C4.5 permite considerar distintos costes relativos de clasificación errónea en la elaboración de los modelos.

Otro aspecto importante a tener en cuenta es que nuestro análisis se ha realizado al margen de factores que probablemente tengan un peso determinante a la hora de predecir la crisis, tales como el tamaño de la empresa o el tipo de negocio en el que opere. Aunque estos factores escaparon del ámbito de nuestro estudio porque estuvimos especialmente interesados en variables de carácter financiero, también podrían incorporarse los factores mencionados como variables predictoras.

Como aplicación práctica del método propuesto nos parece especialmente destacable su utilidad como herramienta de preselección de empresas a investigar más cuidadosamente por parte de la autoridad supervisora. En nuestra opinión, ello facilitaría grandemente la labor de supervisión de las entidades aseguradoras permitiendo que los recursos limitados de la inspección se dirigiesen hacia aquéllas preseleccionadas como potencialmente insolventes, lo que supondría, por tanto, un ahorro de los costes que conlleva la actividad supervisora.

Podrían utilizarse entonces a modo de los *«early warning systems»* empleados por las autoridades supervisoras en algunos países con la finalidad de detectar, con la suficiente antelación, situaciones que pongan de manifiesto problemas financieros en las aseguradoras. Nos referimos fundamentalmente al sistema IRIS *(Insurance Regulatory Information System)* desarrollado en Estados Unidos por la NAIC *(National Association of Insurance Commissioners)* y a su sucesor FAST *(Financial Analysis Tracking System)* creado a comienzos de los años noventa, mediante el cual las aseguradoras son posicionadas en función de su situación frente a una serie de ratios para priorizar a las compañías que requieran un análisis financiero más detallado. De este modo, se pre-

tende identificar a las aseguradoras que tengan, o se prevé que vayan a tener, problemas financieros para facilitar la intervención a tiempo de manera que se evite la insolvencia o se mitiguen al menos los costes asociados a la misma. También se aplican sistemas de supervisión similares en otros países de nuestro entorno como Inglaterra, Italia, Holanda, Canadá, etc.

La utilización más razonable de modelos como nuestros *Modelo 1*, *Modelo 2* y *Modelo 3*, construidos, respectivamente, con los datos del primer, el segundo y el tercero, año previo a la crisis, supondría llevar a cabo lo siguiente:

– Situarnos en el año t para evaluar la situación de una empresa en el año siguiente, $t + 1$. Para ello tomaremos las cuentas anuales de la empresa correspondientes a este ejercicio t, calcularemos los ratios contenidos en el *Modelo 1* y aplicaremos el árbol o las reglas de decisión representativos de dicho *Modelo 1*. Si la empresa resultase ser catalogada como fracasada, pasaría a ser inspeccionada. Si así no fuere, es decir, si la empresa fuese clasificada como sana, calcularíamos los ratios reflejados en el *Modelo 2* para ver qué nos indica este modelo con respecto a la situación de la empresa para dentro de dos años, esto es, en $t+2$. Si ésta fuese catalogada como «mala», sería también objeto de atención especial de manera que pudiésemos llegar a evitar una insolvencia potencial. Si, por el contrario, según el *Modelo 2* la empresa fuese «buena», aplicaríamos entonces el *Modelo 3* que nos indicaría si la empresa será sana o fracasada en $t+3$. Y así seguiríamos si hubiésemos calculado modelos para más horizontes de predicción.
– También podríamos situarnos en el año t para evaluar exclusivamente la situación de la empresa para el año siguiente, $t+1$. Obviamente, sólo con tomar las cuentas anuales de este año t, calcular los ratios contenidos en el *Modelo 1* y aplicar dicho modelo sabremos si la empresa será «buena» o «mala» en $t+1$, pero ya que disponemos de más modelos sería interesante aprovechar también la información proporcionada por los mismos. Entonces la predicción anterior podría complementarse tomando las cuen-

tas anuales de los años anteriores t-1 y t-2, calculando a partir de dichas cuentas, respectivamente, los ratios reflejados en el *Modelo 2* y en el *Modelo 3* y aplicando estos modelos. Lógicamente, los tres modelos nos indicarán lo que ocurrirá con la empresa en t+1. De este modo, como los modelos son independientes entre sí será posible obtener distintas predicciones acerca de la situación de la empresa para t+1. Para decidir si la empresa será «buena» o «mala» podríamos atender al principio del «voto por la mayoría», o simplemente con que fuese clasificada como fracasada por un único modelo habríamos de llevar a cabo un examen minucioso de la empresa.

Asimismo, aunque no se haya hecho aquí, cabría plantearse un objetivo más ambicioso consistente en construir un único modelo que incorpore datos de tres años y que clasifique a las empresas en cuatro categorías, las que quiebran al cabo de un año, de dos, de tres y las que no quiebran en ese período de tres años. Un modelo de este tipo sería preferible a una combinación de tres modelos distintos para uno, dos y tres años, respectivamente, construidos independientemente uno de otro pero usados conjuntamente, ya que permitiría minimizar la probabilidad de error global, mientras que los modelos individuales como los nuestros sólo minimizarían su propia probabilidad de error, lo cual no sería lo más adecuado si lo que se pretendiese fuese minimizar la probabilidad de error del modelo global obtenido usando conjuntamente los modelos individuales.

Resulta evidente que han quedado aquí muchas cuestiones por resolver y sin duda nuestro trabajo puede ser complementado y mejorado, si bien nuestra intención no iba más allá de la de explorar las posibilidades de ciertas técnicas de aprendizaje automático poco explotadas en el ámbito económico y nunca utilizadas con el propósito de predecir las crisis empresariales en el sector del seguro.

Por otro lado, no ignoramos que es en épocas de bonanza económica cuando la identificación y el estudio de los factores que llevaron a las empresas a la quiebra en una fase recesiva del ciclo económico serán primordiales si se pretende que el pasado no se repita en el futuro. En

este sentido, pese a que no podemos pretender haber alcanzado resultados determinantes e incuestionables, sí que esperamos con la realización del presente trabajo —uno más dentro de la plétora de trabajos empíricos acerca del fenómeno del fracaso empresarial— haber contribuido en alguna medida, ciertamente modesta, a la clarificación del problema estudiado.

Referencias bibliográficas

ALTMAN, E. I. (1968): «Financial ratios, discriminant analysis and the prediction of the corporate bankruptcy», *The Journal of Finance*, vol. 23, n.º 4, pp. 589-609.

—; HALDEMAN, R. G. y NARAYANAN, P. (1977): «ZETA™ analysis: a new model to identify bankruptcy risk of corporations», *Journal of Banking & Finance*, vol. 1, n.º 1, pp. 29-54.

—y LORIS, B. (1976): «A Financial early warning system for over-the-counter broker-dealers», *The Journal of Finance*, vol. 31, n.º 4, pp. 1.201-1.217.

ARQUES PÉREZ, A. (1997): *La predicción del fracaso empresarial. Aplicación al riesgo crediticio bancario*, Tesis Doctoral, Universidad de Murcia.

BEAVER, W. H. (1966): «Financial ratios as predictors of failure», *Journal of Accounting Research*, vol. 4, Empirical research in accounting: Selected studies 1966, pp. 71-111.

BEAVER, W. H. (1968): «Alternative accounting measures as predictors of failure». *The Accounting Review*, vol. 43, n.º 1, pp. 113-122.

BEST, A. M. COMPANY (1991): *Best's insolvency study, property/casualty insurers 1969-1990*, Special Report, junio 1991.

BONSÓN PONTE, E.; ESCOBAR RODRÍGUEZ, T. y MARTÍN ZAMORA, M. P. (1996): «Sistemas de inducción de árboles de decisión: utilidad en el análisis de crisis bancarias», *Biblioteca Electrónica CiberConta* (http://ciberconta.unizar.es), Zaragoza.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A. y STONE, C. J. (1984): *Classification and regression trees*. Wadsworth International Group, Belmont, California.

BRODLEY, C. E. y UTGOFF, P. E. (1995): «Multivariate decision trees», *Machine Learning*, vol. 19, n.º 1, pp. 45-77.

CIELEN, A.; PEETERS, L. y VANHOOF, K. (2004): «Bankruptcy prediction using a

data envelopment analysis», *European Journal of Operational Research*, vol. 154, n.º 2, pp. 526-532.

DEAKIN, E. B. (1972): «A discriminant analysis of predictors of business failure», *Journal of Accounting Research,* vol. 10, n.º 1, pp. 167-179.

DEL POZO GARCÍA, E. M. (1997): *Modelos de control de la solvencia en seguros no-vida*, Tesis Doctoral, Universidad Complutense de Madrid.

DIRECTIVA 91/674/CEE del Consejo, de 19 de diciembre de 1991, relativa a las cuentas anuales y a las cuentas consolidadas de las empresas de seguros.

DIZDAREVIC, S., LARRAÑAGA, P., PEÑA, J. M., SIERRA, B., GALLEGO, M. J. y LOZANO, J. A. (1999): «Predicción del fracaso empresarial mediante la combinación de clasificadores provenientes de la estadística y el aprendizaje automático», en E. BONSÓN PONTE (ed.): *Tecnologías Inteligentes para la Gestión Empresarial*, RA-MA Editorial, Madrid, pp. 71-113.

EFRON, B. (1982): *The jackknife, the bootstrap and other resampling plans,* Society for Industrial and Applied Mathematics (SIAM), Filadelfia, Pensilvania.

FAYYAD, U. M. e IRANI, K. B. (1992): «On the handling of continuous-valued attributes in decision tree generation», *Machine Learning*, vol. 8, n.º 1, pp. 87-102.

FEIGENBAUM, E. A. (1981): «Expert systems in the 1980s», en A. BOND (ed.): *State-of-the-art report on machine intelligence*, Pergamon-Infotech, Maidenhead, Inglaterra.

—; MCCORDUCK, P. y NII, H. P. (1988). *The rise of the expert company*, Times Books, Nueva York.

FERNÁNDEZ PALACIOS, J. y MAESTRO, J. L. (1991): *Manual de Contabilidad y Análisis financiero de seguros*, Centro de Estudios del Seguro, Madrid.

FRIEDMAN, J. H. (1977): «A recursive partitioning decision rule for non-parametric classification», *IEEE Transactions on Computers*, vol. 26, n.º 4, pp. 404-408.

—; ALTMAN, E. I. y KAO, D. L. (1985): «Introducing recursive partitioning for financial classification: The case of financial distress», *The Journal of Finance*, vol. 40, n.º 1, pp. 269-291.

GABÁS TRIGO, F. (1990): *Técnicas actuales de Análisis Contable*, Instituto de Contabilidad y Auditoría de Cuentas, Madrid.

— (1997): «Predicción de la insolvencia empresarial», en A. CALVO-FLORES SEGURA y D. GARCÍA PÉREZ DE LEMA (eds.): *Predicción de la insolvencia empresarial*, Asociación Española de Contabilidad y Administración de Empresas, Madrid, pp. 11-31.

GARCÍA PÉREZ DE LEMA, D.; CALVO-FLORES SEGURA, A. y ARQUES PÉREZ A. (1997): «Factores discriminantes del

riesgo financiero en la industria manufacturera española», en A. CALVO-FLORES SEGURA y D. GARCÍA PÉREZ DE LEMA (eds.): *Predicción de la insolvencia empresarial,* Asociación Española de Contabilidad y Administración de Empresas, Madrid, pp. 125-156.

GENTRY, J. A.; NEWBOLD, P. y WHITFORD, D. T. (1985): «Classifying bankrupt firms with funds flow components», *Journal of Accounting Research,* vol. 23, n.º 1, pp. 146-160.

—; SHAW, M. J.; TESSMER, A. C. y WHITFORD, D. T. (2002): «Using inductive learning to predict bankruptcy», *Journal of Organizational Computing and Electronic Commerce,* vol. 12, n.º 1, pp. 39-57.

GONZÁLEZ PÉREZ, A. L.; CORREA RODRÍGUEZ, A. y BLÁZQUEZ MÚREZ, J. A. (1999): «Perfil del fracaso empresarial para una muestra de pequeñas y medianas empresas», *X Congreso AECA - Zaragoza «La empresa española ante el siglo XXI» (Comunicaciones)*, Asociación Española de Contabilidad y Administración de Empresas.

GRACE, M. F.; HARRINGTON, S. E. y KLEIN, R. W. (1998): «Risk-based capital and solvency screening in property-liability insurance: Hypotheses and empirical tests», *The Journal of Risk and Insurance,* vol. 65, n.º 2, pp. 213-243.

HAMPTON, J. (1991): *Financial management of insurance companies*, PCG Publishing, Nueva Jersey.

HERNÁNDEZ ORALLO, J.; RAMÍREZ QUINTANA, M. J. y FERRI RAMÍREZ, C. (2004): *Introducción a la minería de datos,* Pearson Prentice Hall, Madrid.

HUNT, E. B.; MARÍN, J. y STONE, P. J. (1966): *Experiments in induction,* Academic Press, Nueva York.

JENG, B.; JENG, Y. M. y LIANG, T. P. (1997): «FILM: a fuzzy inductive learning method for automated knowledge acquisition», *Decision Support Systems,* vol. 21, n.º 2, pp. 61-73.

JIMÉNEZ CARDOSO, S. M.; GARCÍA-AYUSO COVARSÍ, M. y SIERRA MOLINA, G. J. (2000): *Análisis financiero,* Pirámide, Madrid.

KASS, G. V. (1980): «An exploratory technique for investigating large quantities of categorical data», *Applied Statistics,* vol. 29, n.º 2, pp. 119-127.

LAFFARGA BRIONES, J.; MARTÍN MARÍN, J. L. y VÁZQUEZ CUETO, M. J. (1985): «El análisis de la solvencia en las instituciones bancarias: propuesta de una metodología y aplicaciones a la Banca Española», *Esic Market,* 48, abril-junio, pp. 51-73.

— (1987): «Predicción de la crisis bancaria española: la comparación entre el análisis *logit* y el análisis discriminante», *Cuadernos de Investigación Contable,* vol. 1, n.º 1, pp. 103-110.

LEY 50/1980, de 8 de octubre, de Contrato de Seguro.

LEY 8/1987, de 8 de junio, de Regulación de los Planes y Fondos de Pensiones.

LEY 30/1995, de 8 de noviembre, de Ordenación y Supervisión de los Seguros Privados.

LEY 44/2002, de 22 de noviembre, de Medidas de Reforma del Sistema Financiero.

LIZÁRRAGA DALLO, F. (1996): *Modelos multivariantes de previsión del fracaso empresarial: una aplicación a la realidad de la información contable española*, Tesis Doctoral, Universidad Pública de Navarra.

LÓPEZ HERRERA, D.; MORENO ROJAS, J. y RODRÍGUEZ RODRÍGUEZ, P. (1994): «Modelos de previsión del fracaso empresarial: aplicación a entidades de seguros en España», *Esic Market,* 84, abril-junio, pp. 83-125.

LOZANO ARAGÜÉS, R. (1999): *Análisis práctico de la normativa patrimonial de las entidades aseguradoras,* Centro de Estudios del Seguro, Madrid.

MARAIS, M. L.; PATELL, J. M. y WOLFSON, M. A. (1984): «The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications», *Journal of Accounting Research,* vol. 22, Supplement 1984, pp. 87-114.

MARTÍN PEÑA, M. L.; LEGUEY GALÁN, S. y SÁNCHEZ LÓPEZ, J. M. (1999): *Solvencia y estabilidad financiera en la empresa de seguros: Metodología y evaluación empírica mediante análisis multivariante*, Cuadernos de la Fundación Mapfre Estudios, n.º 49, Madrid.

MARTÍN ZAMORA, M. P. (1999): *La solvencia en las cajas rurales provinciales andaluzas (1978-1985)*, Servicio de Publicaciones, Universidad de Huelva.

MARTÍNEZ DE LEJARZA ESPARDUCER, I. (1999): «Previsión del fracaso empresarial mediante redes neuronales: un estudio comparativo con el análisis discriminante», en E. BONSÓN PONTE (ed.): *Tecnologías inteligentes para la gestión empresarial,* RA-MA Editorial, Madrid, pp. 53-70.

MCKEE, T. E. (1995a): «Predicting bankruptcy via induction», *Journal of Information Technology,* 10, pp. 26-36.

— (1995b): «Predicting bankruptcy via an inductive inferencing algorithm: an extension», en G. J. SIERRA y E. BONSÓN (eds.): *Artificial Intelligence in accounting, finance and tax,* Huelva, pp. 87-97.

MENSAH, Y. M. (1983): «The differential bankruptcy predictive ability of specific price level adjustments: Some empirical evidence», *The Accounting Review,* vol. 58, n.º 2, pp. 228-246.

MESSIER, W. F. y HANSEN, J. V. (1988): «Inducing rules for expert system development: an example using default

and bankruptcy data», *Management Science,* vol. 34, n.º 12, pp. 1.403-1.415.

MICHIE, D. (1987): «Current developments in expert systems», en J. R. QUINLAN (ed.): *Applications of expert systems,* Addison-Wesley, Wokingham, Reino Unido, pp. 137-156.

— (1989): «Problems of computer-aided concept formation», en J. R. QUINLAN (ed.): *Applications of expert systems,* Addison-Wesley, Wokingham, Reino Unido, vol. 2, pp. 310-333.

MILLÁN AGUILAR, A. (2000): *Análisis contable de sociedades aseguradoras*, Asociación Española de Contabilidad y Administración de Empresas, Madrid.

MORA ENGUÍDANOS, A. (1994): «Los modelos de predicción del fracaso empresarial: una aplicación empírica del logit», *Revista Española de Financiación y Contabilidad,* vol. 23, n.º 78, pp. 203-233.

MORGAN, J. N. y MESSENGER, R. C. (1973): *THAID: a sequential search program for the analysis of nominal scale dependent variables,* Technical report, Institute for Social Research, University of Michigan, Ann Arbor, Míchigan.

—y SONQUIST, J. A. (1963): «Problems in the analysis of survey data, and a proposal», *Journal of the American Statistical Association,* 58, pp. 415-434.

MURTHY, S. K. (1998): «Automatic construction of decision trees from data: A multi-disciplinary survey», *Data Mining and Knowledge Discovery,* vol. 2, n.º 4, pp. 345-389.

ORDEN de 30 de julio de 1981 por la que se aprueban las normas de adaptación del Plan General de Contabilidad a las Entidades de Seguros, Reaseguros y Capitalización.

PEÑA, D. (2002): *Análisis de datos multivariantes*, McGraw-Hill, Madrid.

PINA MARTÍNEZ, V. (1989): «La información contable en la predicción de la crisis bancaria 1977-1985», *Revista Española de Financiación y Contabilidad,* vol. 18, n.º 58, pp. 309-338.

QUINLAN, J. R. (1979): «Discovering rules by induction from large collections of examples», en D. MICHIE (ed.): *Expert systems in the micro electronic age*, Edinburgh University Press, Edimburgo, Reino Unido, pp. 168-201.

— (1983): «Learning efficient classification procedures and their application to chess endgames», en R. S. MICHALSKI; J. G. CARBONELL y T. M. MITCHELL (eds.): *Machine Learning: An Artificial Intelligence approach*, Tioga Publishing Company, Palo Alto, California, pp. 463-482.

— (1986a): «The effect of noise on concept learning», en R. S. MICHALSKI; J. G. CARBONELL y T. M. MITCHELL (eds.): *Machine Learning: An Artificial Intelligence approach,* Morgan Kaufmann, San Mateo, California, vol. 2, pp. 149-166.

QUINLAN, J. R. (1986b): «Induction of decision trees», *Machine Learning,* vol. 1, n.º 1, pp. 81-106.

— (1987a): «Generating production rules from decision trees», en *Proceedings of the Tenth International Joint Conference on Artificial Intelligence,* Morgan Kaufmann, San Mateo, California, pp. 304-307.

— (1987b): «Simplifying decision trees», *International Journal of Man-Machine Studies,* vol. 27, n.º 3, pp. 221-234.

— (1988): «Decision trees and multivalued attributes», en J. E. HAYES, D. MICHIE y J. RICHARDS (eds.): *Machine Intelligence 11 - Towards an automated logic of human thought,* Clarendon Press, Oxford, Reino Unido, pp. 305-318.

— (1989): «Unknown attribute values in induction», *Proceedings of the Sixth International Machine Learning Workshop,* Morgan Kaufmann, San Mateo, California, pp. 164-168.

— (1990): «Decision trees and decision-making», *IEEE Transactions on systems, man and cybernetics,* vol. 20, n.º 2, pp. 339-346.

— (1991): «Improved estimates for the accuracy of small disjuncts», *Machine Learning,* vol. 6, n.º 1, pp. 93-98.

— (1993): *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California.

QUINLAN, J. R. (1996): «Improved use of continuous attributes in C4.5», *Journal of Artificial Intelligence Research,* 4, pp. 77-90.

—; COMPTON, P. J.; HORN, K. A. Y LAZARUS, L. A. (1987): «Inductive knowledge acquisition: A case study», en J. R. QUINLAN (ed.): *Applications of Expert Systems*, Addison-Wesley, Wokingham, Reino Unido, pp. 157-173.

—y RIVEST, R. L. (1989): «Inferring decision trees using the Minimum Description Length Principle», *Information and Computation*, vol. 80, n.º 3, pp. 227-248.

R DEVELOPMENT CORE TEAM (2005): *R: A language and environment for statistical computing,* R Foundation for Statistical Computing, Viena, Austria (http://www.R-project.org).

REAL DECRETO 2014/1997, de 26 de diciembre, por el que se aprueba el Plan de Contabilidad de las entidades aseguradoras y normas para la formulación de las cuentas de los grupos de entidades aseguradoras.

REAL DECRETO 2486/1998, de 20 de noviembre, por el que se aprueba el Reglamento de Ordenación y Supervisión de los Seguros Privados.

REAL DECRETO 297/2004, de 20 de febrero, por el que se modifica el Reglamento de Ordenación y Supervisión de los Seguros Privados, aprobado por el Real Decreto 2486/1998, de 20 de Noviembre.

REAL DECRETO 298/2004, de 20 de febrero, por el que se modifica el Plan de Contabilidad de las entidades aseguradoras y normas para la formulación de las cuentas de los grupos de entidades aseguradoras, aprobado por el Real Decreto 2014/1997, de 26 de diciembre.

REAL DECRETO LEGISLATIVO 6/2004, de 29 de octubre, por el que se aprueba el Texto Refundido de la Ley de Ordenación y Supervisión de los Seguros Privados.

REAL DECRETO LEGISLATIVO 7/2004, de 29 de octubre, por el que se aprueba el Texto Refundido del Estatuto Legal del Consorcio de Compensación de Seguros.

REAL DECRETO-LEY 10/1984, de 11 de julio, por el que se establecen medidas urgentes para el saneamiento del sector de seguros privados y para el reforzamiento del organismo de control.

REZA, F. M. (1994): *An introduction to Information Theory*, Dover Publications, Nueva York.

RISSANEN, J. (1983): «A universal prior for integers and estimation by minimum description length», *Annals of Statistics*, vol. 11, n.º 2, pp. 416-431.

RODRÍGUEZ ACEBES, M. C. (1990): *La predicción de las crisis empresariales. Modelos para el sector de seguros*, Secretariado de publicaciones, Universidad de Valladolid, Serie Economía, n.º 14.

SALCEDO SANZ, S.; FERNÁNDEZ VILLACAÑAS, J. L.; SEGOVIA VARGAS, M. J. Y BOUSOÑO CALZÓN, C. (2005): «Genetic programming for the prediction of insolvency in non-life insurance companies», *Computers & Operations Research,* vol. 32, n.º 4, pp. 749-765.

SANCHIS ARELLANO, A. (2000): *Una aplicación del Análisis Discriminante a la previsión de la insolvencia en las empresas españolas de seguros no vida,* Tesis Doctoral, Universidad Complutense de Madrid.

—; GIL, J. A. y HERAS MARTÍNEZ, A. (2003): «El Análisis Discriminante en la previsión de la insolvencia en las empresas de seguros de no vida», *Revista Española de Financiación y Contabilidad,* vol. 32, n.º 116, pp. 183-233.

SEGOVIA VARGAS, M. J. (2003): *Predicción de crisis empresariales en seguros no vida mediante la metodología* Rough Set, Tesis Doctoral, Universidad Complutense de Madrid.

—; SALCEDO SANZ, S. y BOUSOÑO CALZÓN, C. (2004): «Prediction of insolvency in non-life insurance companies using support vector machines, genetic algorithms and simulated annealing», *Fuzzy Economic Review,* vol. 9, n.º 1, pp. 79-94.

SERRANO CINCA, C. Y MARTÍN DEL BRÍO, B. (1993): «Predicción de la quiebra bancaria mediante el empleo de redes neuronales artificiales», *Revista Es-*

pañola de Financiación y Contabilidad, vol. 22, n.º 74, pp. 153-176.

STEWART, B. D. (1987): «Profit cycles in property-liability insurance», en E. D. RANDALL (ed.): *Issues in insurance,* American Institute for Property and Liability Underwriters, Malvern, Pensilvania, vol. 2, pp. 111-174.

SWAIN, P. y HAUSKA, H. (1977): «The decision tree classifier: Design and potential», *IEEE Transactions on Geoscience Electronics*, vol. GE-15, n.º 3, pp. 142-147.

TAM, K. Y. y KIANG, M. Y. (1992): «Managerial applications of neural networks: the case of bank failure predictions», *Management Science*, vol. 38, n.º 7, pp. 926-947.

TROYANSKAYA, O.; CANTOR, M.; SHERLOCK, G.; BROWN, P.; HASTIE, T.; TIBSHIRANI, R.; BOTSTEIN, D. y ALTMAN, R. B. (2001): «Missing value estimation methods for DNA microarrays», *Bioinformatics*, vol. 17, n.º 6, pp. 520-525.

WITTEN, I. H. y FRANK, E. (2000): *Data mining: Practical Machine Learning tools and techniques with java implementations*, Morgan Kaufmann, San Francisco, California.

COLECCIÓN LÍNEA 3000

TOXINOLOGÍA CLÍNICA, ALIMENTARÍA Y AMBIENTAL
Miguel Capó Martí
ISBN: 978-84-7491-879-3; 176 págs.; 12,00 €

EL EROTISMO EN LA POESÍA DE ADÚLTEROS Y CORNUDOS EN EL SIGLO DE ORO
Félix Cantizano Pérez
ISBN: 978-84-7491-854-0; 128 págs.; 12,00 €

TEORÍA KANTIANA DE LA VOLUNTAD. ESTUDIO EN ANTROPOLOGÍA EN SENTIDO PRÁGMATICO
Alejandro García Mayo
Próxima publicación en libro electrónico

PREDICCIÓN DE CRISIS EMPRESARIALES EN SEGUROS NO VIDA, MEDIANTE ÁRBOLES DE DECISIÓN Y REGLAS DE CLASIFICACIÓN
Zuleyka Díaz Martínez
Próxima publicación en libro electrónico

COLECCIÓN CLÁSICOS BREVES

SOBRE EL CONCEPTO DE VERDAD
Franz Brentano
ISBN: 978-84-7491-804-5; 48 págs.; 3,00 €

LA TIERRA NO SE MUEVE
Edmund Husserl
ISBN: 978-84-7491-803-8; 64 págs.; 3,00 €

PRUDENCIA, MORALIDAD Y EL DILEMA DEL PRISIONERO
Derek Parfit
ISBN: 978-84-7491-853-3; 72 págs.; 3,00 €

COLECCIÓN FORO COMPLUTENSE

¿ES POSIBLE ACABAR CON LA POBREZA?
Muhammad Yunus
ISBN: 978-84-7491-802-1; 48 págs.; 3,00 €

NO PIENSES EN UN ELEFANTE. LENGUAJE Y DEBATE POLÍTICO
George Lakoff
ISBN: 978-84-7491-813-7; 176 págs.: 10,00 €

EL ISLAM EN EUROPA. ¿UNA RELIGIÓN MÁS O UNA CULTURA DIFERENTE?
Olivier Roy
ISBN: 978-84-7491-806-9; 48 págs.; 3,00 €

DIÁLOGO DE CULTURAS E IDENTIDADES
Sami Naïr
ISBN: 978-84-7491-805-2; 48 págs.; 3,00 €

TERROR SAGRADO
Terry Eagleton
ISBN: 978-84-7491-848-9; 56 págs.; 3,00 €

MULTIMEGAMUCHOGLOBALIZACIÓN
Jose Luis Sanpedro y Carlos Berzosa
Próxima publicación

OTROS TÍTULOS EDITORIAL COMPLUTENSE

DICCIONARIO DE RELACIONES INTERCULTURALES. DIVERSIDAD Y GLOBALIZACIÓN
Ascensión Barañano, José L. García, María Cátedra y Marie Jose Devillard (coords.)
ISBN: 978-84-7491-814-4; 448 págs.; 28,00 €

VENUS VENERADA. TRADICIONES ERÓTICAS DE LA LITERATURA ESPAÑOLA
J.Ignacio Díez y Adrienne Martín (eds.)
Colección Académica; ISBN: 978-84-7491-791-8; 280 págs.; 15,00 €

VENUS VENERADA II. LITERATURA ERÓTICA Y MODERNIDAD EN ESPAÑA
Adrienne L. Martín y J. Ignacio Díez (eds.)
Colección Académica; ISBN: 978-84-7491-839-7; 344 págs.; 15,00 €

LA DESTRUCCIÓN DE LA CIENCIA EN ESPAÑA. DEPURACIÓN UNIVERSITARIA EN EL FRANQUISMO
Luis Enrique Otero (ed.)
ISBN: 978-84-7491-808-3; 384 págs.; 20,00 €

DICCIONARIOS OXFORD/COMPLUTENSE
De Arte del siglo XX, Historia Universal del siglo XX, Ciencias de la Tierra, Enfermería, Física, Matemáticas, Medicina, Química...

LA GUERRA EN EL IMPERIO AZTECA. EXPANSIÓN, IDEOLOGÍA Y ARTE
Isabel Bueno Bravo
Colección la Mirada de la Historia
Próxima publicación

PENSAR EL FINAL: LA EUTANASIA. ÉTICAS EN CONFLICTO
Luis Montiel Llorente y María García Alonso (eds.)
Colección Pensar nuestro tiempo; ISBN: 978-84-7491-842-7-3; 200 págs.; 15,00 €