Un journal de asfae chile

mbr Multidisciplinary Business Review

LISTAS DE DECISIÓN APLICADAS A LA PREDICCIÓN DE LA QUIEBRA EN EL SECTOR DEL SEGURO NO VIDA*

DECISION LISTS APPLIED TO THE BANKRUPTCY PREDICTION IN NON-LIFE INSURANCE SECTOR*

Zuleyka Díaz Martínez[a] • José A. Gil Fana[b] • Eva M. Pozo García[c]

Clasificación: Trabajo empírico - investigación
Recibido: Noviembre 2009 / Aceptado: Diciembre 2009
*Este trabajo ha sido financiado por Banco Santander Central Hispano y Universidad Complutense de Madrid a través del proyecto de investigación ref. SANTANDER-UCM PR34/07-15788.

Resumen

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 estas razones planteamos este trabajo, con el que pretendemos desarrollar un modelo fácilmente interpretable a través de la aplicación de un método de implementación sencilla procedente del campo de la Inteligencia Artificial, el algoritmo de inducción de listas de decisión PART, con el propósito de comprobar el grado de aplicabilidad de dicho algoritmo a la valoración de la solvencia de las empresas de seguros. Asimismo, compararemos los resultados alcanzados con los que se obtienen mediante la aplicación de Regresión Logística.

Palabras clave: Quiebra, sector del seguro, Inteligencia Artificial, listas de decisión.

Abstract

Prediction of insurance companies insolvency has arisen as an important problem in the field of financial research, in order to protect both society and customers and minimize the costs associated with this problem.

The development and application of new criteria for a better harnessing of the financial-accounting information furnished by insurance companies is a key issue to handle this problem. In line with this issue, this paper aims to examine the applicability of a technique coming from the Artificial Intelligence to the prediction of insolvency in the insurance sector, a learning algorithm for decision lists called PART. We also compare this method with a classical statistical approach, Logistic Regression. PART presents the advantage of being easy to implement as well as providing results of simple interpretation whilst avoiding some of the drawbacks of the conventional statistical techniques.

Keywords: Bankruptcy, insurance sector, Artificial Intelligence, decision lists.

a Universidad Complutense de Madrid. Facultad de Ciencias Económicas y Empresariales. Email: zuleyka@ccee.ucm.es.

b Universidad Complutense de Madrid. Facultad de Ciencias Económicas y Empresariales. Email: jagilfan@ccee.ucm.es.

c Universidad Complutense de Madrid. Facultad de Ciencias Económicas y Empresariales. Email: epozo@ccee.ucm.es.

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 caracterizar 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 se ha sugerido recientemente 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 decisor 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 listas de decisión PART (Frank y Witten, 1998). Concretamente, el propósito de este trabajo es comprobar el grado de aplicabilidad del algoritmo PART a la valoración de la solvencia de las empresas de seguros, siendo el fin último de nuestra investigación el desarrollar un conjunto de modelos sencillos basados en ratios financieros en forma de listas 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 la sección 1 se exponen brevemente los fundamentos teóricos del algoritmo PART. La sección 2 se refiere a todos los aspectos relativos a la selección de datos y variables que intervienen en nuestro estudio empírico. En la sección 3 se presentan los resultados obtenidos con el algoritmo PART y su comparación con la Regresión Logística. Finalmente, en la sección 4 exponemos nuestras conclusiones.

1. El algoritmo PART

Como ya hemos mencionado, los algoritmos de Aprendizaje Automático construyen modelos automáticamente que describen la estructura subyacente en un conjunto de datos, esto es, inducen un modelo u output a partir de un conjunto de observaciones o input. 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, el recíproco no es necesariamente cierto: algunos modelos son diseñados exclusivamente para la predicción y su habilidad en esta faceta no implica su utilidad para el análisis. En muchas aplicaciones 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 dominio. 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 –.

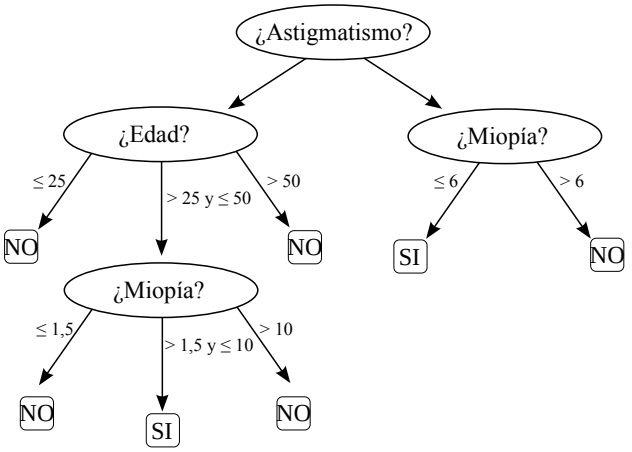
Dos de los enfoques más fructíferos y ampliamente utilizados de Aprendizaje Automático son los constituidos por los Árboles de Decisión y las Reglas de Clasificación, que pese a sus diferencias de carácter formal guardan una estrecha relación que hace que puedan ser consideradas como distintas variantes de una metodología común. 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 donde son importantes tanto la capacidad predictiva como la utilidad para el análisis, esto es, la predicción y la explicación. En nuestro trabajo utilizaremos precisamente una de dichas técnicas, el algoritmo de generación de reglas PART, que se apoya en el conocido algoritmo de inducción de árboles de decisión C4.5.

Los árboles de decisión son un modo de representación de la regularidad subyacente en los datos en forma de un conjunto de condiciones excluyentes y exhaustivas organizadas en una estructura jerárquica arborescente compuesta 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 -en nuestro caso, una empresa caracterizada por un conjunto de ratios financieros-, 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. Así, por ejemplo, la siguiente figura (**Figura 1.1**) muestra cómo podría determinarse mediante un árbol de decisión la recomendación de cirugía refractiva a pacientes miopes en un hospital, 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".

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



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

En cuanto al mecanismo de generación de un árbol, existe una gran diversidad de ellos pero todos se basan en utilizar un conjunto de casos de entrenamiento sobre el que se van haciendo particiones recursivas (el conjunto se divide sucesivamente dotándole de una estructura ramificada) de acuerdo con ciertas reglas que se seleccionan de manera que se minimice una "función de impureza" que mida el grado en que los distintos subconjuntos generados son más o menos puros, es decir, sus elementos son más o menos homogéneos (entendida la homogeneidad en el sentido de pertenencia a la misma clase).

Aunque ha sido propuesta una gran variedad de funciones de impureza, existen algunas especialmente relevantes y cuyo uso está ampliamente extendido. Éste es el caso del *índice de Gini*, empleado en el sistema CART -*Classification and Regression Trees*- (Breiman et al., 1984), y las medidas basadas en la entropía, como la "*ganancia de información*" o el "*ratio de ganancia*", utilizadas en C4.5. Este último 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 ochenta y principios de los noventa (Quinlan, 1993) como descendiente de un primer programa clasificador que fue denominado ID3 (Quinlan, 1979, 1983, 1986). El algoritmo C4.5 se basa en la entropía de una variable aleatoria (que mide la incertidumbre asociada a dicha variable) y la información mutua entre variables distintas (que indica la reducción de incertidumbre con respecto a una de ellas cuando conocemos el valor de la otra u otras) para desarrollar un conjunto de reglas, esencialmente heurísticas pero notablemente ingeniosas y de gran eficacia, que permiten construir árboles de decisión a partir de conjuntos de casos de prueba. El algoritmo admite tanto variables continuas como discretas (categóricas) e incorpora otras características adicionales que le dotan de gran potencia y flexibilidad como es, por ejemplo, su capacidad para manejar valores faltantes (*missing values*). Así, los casos que presentan un *missing value* para algún atributo son fraccionados cuando llegan a un nodo del árbol en el cual se toma una decisión de acuerdo con los valores del atributo faltante. Al ser tal valor desconocido para el caso en cuestión, este caso se divide o reparte entre las distintas ramas que salen del nodo de acuerdo con la proporción en la que lo hacen los casos para los cuales el valor del atributo es conocido. Esto hace que surjan valores fraccionarios para el número de casos que llegan a los nodos y hojas del árbol. De la misma manera, cuando el árbol es utilizado para clasificar un nuevo caso y alguno de sus atributos tiene un valor desconocido, el fraccionamiento anterior da lugar a que lo que se obtenga no sea una clasificación determinista o unívoca, sino una distribución de probabilidad sobre las clases a las que eventualmente el caso pueda pertenecer para finalmente asignarlo a aquella clase para la cual la probabilidad de pertenencia sea máxima.

Generalmente, los métodos recursivos de construcción de árboles de decisión conducirán 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 ele-

mentos, 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.

Así, el algoritmo C4.5 implementa un método de poda 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, C4.5 considera que la existencia de una hoja que cubre N casos clasificando incorrectamente E de ellos puede ser interpretada 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. Entonces, para una hoja que cubra N casos, el número de errores previstos será $N \times p_e$. Similarmente, el número de errores previstos asociados con un subárbol será la suma de los de cada una de sus ramas, y los 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.

Por otra parte, aunque los árboles de decisión representan el conocimiento de manera muy sencilla, su inteligibilidad disminuye conforme aumenta su tamaño. Un conjunto de reglas de decisión de la forma si (condiciones) - *entonces* (decisión) es un mecanismo alternativo 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.

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 quizás una distribución de probabilidad sobre las clases.

Los algoritmos de inducción de árboles de decisión se basan en un enfoque denominado "divide y vencerás" (*divide-and-conquer*): trabajan "de arriba a abajo" - por ello se emplea con frecuencia el acrónimo TDIDT (*Top-Down Induction on Decision Trees*) para hacer referencia a la familia de algoritmos de construcción de árboles de decisión -, 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. No obstante, la conversión de un árbol en reglas no es tan trivial cuando se trata de producir reglas eficaces.

Un enfoque alternativo de construcción de reglas consiste en tomar cada una de las clases del problema por turnos y buscar un modo de cubrir todas las instancias de la clase considerada, excluyendo al mismo tiempo las instancias que no pertenezcan a esa clase. Este enfoque se denomina de *cobertura*, porque en cada nivel se identifica una regla que "cubre" algunas de las instancias. Mientras que las particiones de un árbol de decisión tienen en cuenta todas las clases del problema, intentando maximizar la pureza de la partición, estos métodos de generación de reglas se concentran cada vez en una sola clase, desatendiendo a las otras clases. Son técnicas que siguen una estrategia "separa y vencerás" (*separate-and-conquer*), porque identifican una regla que cubre instancias de la clase deseada (y excluye las de otras clases), separan dichas instancias, y continúan procesando las restantes.

Los algoritmos de cobertura operan añadiendo condiciones a la regla que se esté construyendo mientras vayan cubriendo ejemplos de una manera consistente, siempre con el objetivo de crear una regla con la máxima precisión. En contraste, los algoritmos de partición operan añadiendo condiciones al árbol que estén construyendo, siempre con el objetivo de maximizar la separación entre

las clases.

Si en los árboles de decisión generados mediante algoritmos de partición o *divide-and-conquer* las condiciones son excluyentes y exhaustivas, ya sean representados en forma de árbol o en forma de reglas, esto no es así para los conjuntos de reglas generados mediante algoritmos de cobertura o *separate-and-conquer*, pues en este caso varias reglas podrían ser aplicables para la misma instancia. Además, pueden existir reglas contradictorias para algunos ejemplos. Esto puede resolverse dando un orden a las reglas (obteniéndose entonces las denominadas *listas de decisión*) o ponderando las predicciones diversas.

Las listas de decisión pueden considerarse reglas SI – ENTONCES extendidas y tienen la forma:

si … entonces … ; si no:
si … entonces … ; si no:
si … entonces … ; si no:

La estructura ordenada de las listas de decisión elimina el solapamiento entre las reglas al que se suelen achacar las ineficiencias de algunos algoritmos de inducción de reglas (Berzal Galiano, 2002). Con una lista de decisión, al clasificar un ejemplo se va emparejando dicho ejemplo con cada una de las reglas de la lista hasta que se verifica el antecedente de una de ellas y, entonces, se le asigna al ejemplo la clase que aparece en el consecuente de la regla activada. Por si se diese el caso de que no se verificase ninguna de las reglas de la lista de decisión, usualmente se añade al final de la lista una regla por defecto con antecedente vacío que corresponde a la clase más común de los ejemplos del conjunto de entrenamiento no cubiertos por las reglas seleccionadas (o, en su defecto, la clase más común en el conjunto de entrenamiento completo). De este modo, nunca habrá conflictos entre las reglas.

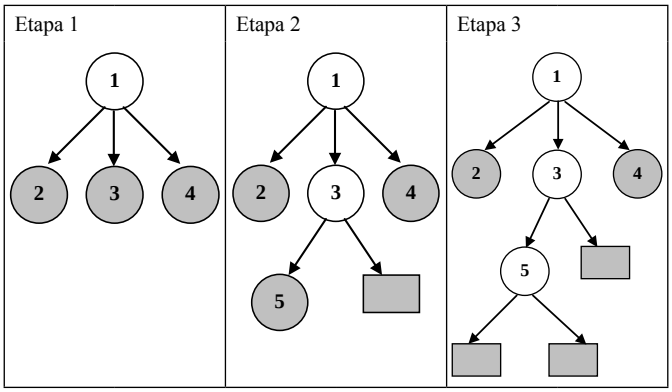
El algoritmo PART de aprendizaje de reglas basado en árboles de decisión parciales (Frank y Witten, 1998) representa un enfoque alternativo híbrido para la inducción de listas de decisión, híbrido porque combina la estrategia *divide-and-conquer* de aprendizaje de árboles de decisión con la estrategia *separate-and-conquer* de aprendizaje de reglas. Adopta la estrategia *separate-and-conquer* en el sentido de que construye una regla, elimina las instancias que ésta cubre y continúa creando reglas recursivamente para las instancias que permanecen hasta que no quede ninguna. Sin embargo, difiere del enfoque estándar en el modo en que se crea cada regla. En esencia, para crear una regla, se construye un árbol de decisión podado a partir del conjunto activo de instancias, la hoja de éste con mayor cobertura se convierte en una regla, y se desecha el árbol.

Aunque el hecho de construir repetidamente árboles de decisión para simplemente descartar la mayoría de ellos pueda resultar un tanto extraño, en verdad resulta que el empleo de un árbol podado para obtener una regla en vez de construirla incrementalmente añadiendo conjunciones evita la tendencia a la "sobrepoda", un problema característico de los algoritmos básicos *separate-and-conquer* de aprendizaje de reglas. La utilización de la metodología *separate-and-conquer* en conjunción con árboles de decisión añade flexibilidad y velocidad. Construir un árbol de decisión completo para obtener una única regla supondría un enorme despilfarro de recursos, pero el proceso puede ser significativamente acelerado sin sacrificio de las ventajas mencionadas de la manera implementada en PART: la idea clave es construir un árbol de decisión parcial en vez de uno completo. Un árbol de decisión parcial contiene algunas ramas que representan subárboles no definidos. Para generar tal árbol parcial, se integran las operaciones de construcción y poda con el objetivo de encontrar un subárbol "estable" que no pueda simplificarse más. Una vez hallado este subárbol, la construcción del árbol cesa y dicho subárbol se convierte en una regla.

El proceso es el siguiente: en primer lugar se escoge una pregunta del mismo modo que en C4.5 para dividir el conjunto de instancias de acuerdo con ella. A continuación, los subconjuntos resultantes se expanden en orden creciente de acuerdo con su entropía, empezando con el de menor entropía, debido a que es más probable que la expansión de los subconjuntos de baja entropía finalice rápidamente y dé lugar a subárboles de pequeño tamaño y por lo tanto a reglas más generales. La expansión se va realizando recursivamente, pero tan pronto como aparezca un nodo interno cuyos hijos ya se hayan expandido en hojas, se comprueba si dicho nodo interno puede ser sustituido por una única hoja, esto es, se intenta "podar" ese subárbol, y la decisión acerca de esta poda se toma de la misma manera que en C4.5. Si el reemplazo se lleva a cabo, se vuelve hacia atrás a explorar los nodos hermanos del nodo reemplazado. Sin embargo, si durante la exploración se encuentra un nodo cuyos hijos no sean todos hojas –lo que sucederá tan pronto como el potencial reemplazo de un subárbol no se lleve a cabo– los subconjuntos restantes ya no se explorarán y, por tanto, los subárboles correspondientes no serán definidos, deteniéndose automáticamente la generación del árbol.

La siguiente figura muestra un ejemplo del proceso:

Figura 1.2. Ejemplo de construcción de un árbol parcial.



Fuente: Frank y Witten (1998).

Desde la etapa 1 hasta la 3, se lleva a cabo la construcción del árbol recursivamente del modo usual, pero escogiendo para la expansión el nodo con la entropía más baja, en este ejemplo, el nodo 3 entre las etapas 1 y 3. El resto de nodos circulares todavía no son expandidos. Los nodos rectangulares representan hojas. Entre las etapas 2 y 3, el nodo rectangular tendrá una entropía más baja que su hermano, el nodo 5, pero no puede ser expandido porque ya es una hoja. Entonces se vuelve hacia atrás y el nodo 5 resulta elegido para su expansión. Cuando se alcanza la tercera etapa existe un nodo cuyos hijos son todos hojas, el nodo 5, y esto desencadena el proceso de poda. Se plantea la posibilidad de reemplazar este subárbol, y se acepta tal reemplazo, lo que conduce a la etapa 4. Ahora se considera el nodo 3 para su reemplazo, y de nuevo es aceptado. El retroceso continúa y ahora resulta que el nodo 4 tiene una entropía más baja que el 2; entonces el nodo 4 se expande en 2 hojas. Se estudia la posibilidad de su reemplazo, y supongamos que el nodo 4 resulta no ser reemplazado. En este punto, el proceso finalizaría, habiéndose obtenido el árbol parcial de 3 hojas de la etapa 5.

Una vez construido un árbol parcial, se extraerá una única regla a partir de él. Cada una de sus hojas se corresponde con una regla posible, y se escogerá la que cubra el mayor número de instancias, puesto que proporcionará la regla más general.

Si se está construyendo un árbol parcial y existen instancias con valor desconocido para alguno de los atributos implicados, su tratamiento será similar al empleado en el algoritmo C4.5. Cuando la lista de decisión obtenida vaya a ser utilizada para clasificar una nueva instancia con atributos desconocidos, se generará una distribución de probabilidad sobre las clases correspondientes a las distintas reglas que le puedan ser aplicadas. La fracción del caso que se asigna a cada una de estas reglas vendrá dada por el porcentaje de casos de entrenamiento que llegando a la regla son cubiertos por ella. Finalmente, la clase más probable de acuerdo con la distribución de probabilidad así obtenida será la que se asigne a la nueva instancia que se está clasificando.

De acuerdo con los experimentos realizados por sus creadores, el algoritmo PART produce con gran rapidez conjuntos de reglas tan o más precisos que otros métodos rápidos de inducción de reglas. Pero su principal ventaja sobre otras técnicas no es el rendimiento sino la simplicidad, y ello se consigue combinando el método de inducción *top-down* de árboles de decisión con la estrategia *separate-and-conquer* de aprendizaje de reglas. Estas razones son las que nos han conducido a decantarnos por dicho algoritmo para la realización de nuestro trabajo.

2. Selección de datos y variables

La muestra de empresas que hemos utilizado en nuestro análisis es la seleccionada por Sanchis Arellano *et al.* (2003) para la aplicación del Análisis Discriminante a la predicción de la insolvencia en empresas españolas de seguros no vida.

La muestra abarca datos del periodo 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 aquéllas que fueron intervenidas por la Comisión Liquidadora de Entidades Aseguradoras (CLEA) y la otra por 36 empresas no fracasadas- que para los mismos periodos se mantenían en funcionamiento.

La muestra comprende entonces un total de 72 empresas españolas de seguros no vida, todas ellas Sociedades Anónimas, por entender 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); Martínez de Lejarza Esparducer (1999).

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 periodo 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 periodos 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.

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.

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. En la siguiente tabla (**Tabla 2**) se exponen los 25 ratios seleccionados como variables independientes a introducir en los modelos:

Tabla II: Ratios empleados

Ratio	Definición
R1	(Inversiones + Tesorería) / (Provisiones técnicas + Depósitos recibidos)
R2	Neto patrimonial / (Inmovilizado + Créditos + Ajustes periodificación (del activo) – Deudas – Provisión para riesgos y gastos – Ajustes periodificación (del pasivo))
R3	Activo circulante / Pasivo circulante
R4	Activo real / Pasivo exigible
R5	Pasivo exigible / Neto
R6	Provisiones técnicas seguro directo / Total primas seguro directo
R7	Provisiones técnicas negocio neto / Total primas negocio neto
R8	Provisiones técnicas seguro directo / Fondos propios
R9	Provisiones técnicas negocio neto / Fondos propios
R10	Total primas seguro directo / Fondos propios
R11	Total primas negocio neto / Fondos propios
R12	Gastos técnicos seguro directo / Fondos propios
R13	Gastos técnicos negocio neto / Fondos propios
R14	Gastos técnicos seguro directo / (Fondos propios + Provisiones técnicas)
R15	Gastos técnicos negocio neto / (Fondos propios + Provisiones técnicas netas)
R16	Comisiones sobre el reaseguro cedido / Fondos propios
R17	Gastos técnicos seguro directo / Primas adquiridas seguro directo

R18	Gastos técnicos negocio neto / Primas adquiridas negocio neto
R19	Gastos de gestión netos / Total primas negocio neto
R20	$\frac{\text{Gastos técnicos seguro directo}}{\text{Primas adquiridas seguro directo}} + \frac{\text{Gastos de gestión}}{\text{Total primas seguro directo}}$
R21	$\frac{\text{Gastos técnicos negocio neto}}{\text{Primas adquiridas negocio neto}} + \frac{\text{Gastos de gestión netos}}{\text{Total primas negocio neto}}$
R22	Ingresos financieros / (Tesorería + Inversiones)
R23	Beneficio antes de impuestos / Fondos propios
R24	Beneficio antes de impuestos / Pasivo total
R25	Cash-flow / Pasivo total

3. Listas de decisión aplicadas 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 listas de decisión PART 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 la sección precedente.

Como ya se ha mencionado, una vez tomada la muestra nos situamos en periodos 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 la sección 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 determinístico, 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 sería 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 las listas de decisión construidas con el algoritmo PART de cara a predecir el fracaso empresarial en las empresas de seguros, sin embargo 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 de la presente sección 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 las listas de decisión PART hemos utilizado el paquete gratuito de minería de datos *WEKA* desarrollado en la Universidad de Waikato en Nueva Zelanda por los propios autores del algoritmo PART (Witten y Frank, 2005), y para la aplicación de la Regresión Logística hemos empleado el software *R 2.1.0* distribuido gratuitamente por CRAN Foundation (R Development Core Team, 2005).

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 resultan 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, PART 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.

3.2. Resultados

Durante el proceso de generación de las listas de decisión hemos tratado de impedir en la medida de lo posible la construcción de listas muy complejas y excesivamente ajustadas 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.

Tomando los datos del primer año previo a la quiebra, para evitar en la medida de lo posible el problema al que acabamos de referirnos hemos exigido que cada regla de la lista cubra al menos 8 empresas, de manera que no se generen reglas para casos muy concretos que den lugar a una lista excesivamente compleja. De este modo, la lista de decisión obtenida para este primer año anterior a la quiebra ha resultado ser bastante sencilla, como puede verse a continuación:

Figura 3.1. Lista de decisión *Modelo 1*

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

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

PART decision list
------------------

R3  <= 1.401888 AND
R6  <= 0.964601: mala (18.37)

R6  > 1.123405: buena (12.0)

R20  <= 1.181138 AND
R4  <= 2.324565 AND
R4  > 1.863728: buena (11.0)

R6  > 0.373819 AND
R4  > 1.463728: buena (11.0)

R23  <= -0.013076: mala (10.63/2.0)

: mala (9.0)

Number of Rules  :      6
```

En la figura anterior puede observarse que la lista de decisión se compone de seis reglas, la última de ellas la regla por defecto. La primera de las reglas consta de dos condiciones en su antecedente. Cuando se trate de clasificar una nueva empresa, si se cumplen las dos condiciones dicha empresa será clasificada como "mala" (fracasada). Si esta regla no fuese aplicable, se pasaría a la siguiente, y si se verificase la única condición de esta segunda regla la empresa sería clasificada como "buena" (sana), y así sucesivamente. Por último, si no se verificase ninguna de las reglas secuenciales, a la empresa se le asignaría la clase por defecto ("mala").

Al final de cada una de las reglas de la lista se observan entre paréntesis unos valores *n* o n/*m*: *n* representa el número de empresas del conjunto de entrenamiento que se clasifican de acuerdo con esa regla y *m* el número de errores cometidos por la misma, es decir, el número de empresas clasificadas incorrectamente. Cuando, como en este caso, alguno de los valores sea fraccionario, será por causa del tratamiento de los *missing values* incorporado en PART que se ha comentado previamente. Como

se observa en la lista, sólo dos empresas del conjunto de entrenamiento resultan ser mal clasificadas, puesto que son asignadas a la clase "mala" siendo en realidad empresas no fracasadas. Esto supone un porcentaje de acierto del 100% con las empresas "malas" y del 94.4% con las "buenas", o el 97.2% en global. Pero este excelente resultado no es representativo de la capacidad de generalización del modelo. Éste podría ser muy específico, muy ajustado a esos datos concretos, y comportarse mal en la clasificación de nuevos elementos. Para comprobarlo, llevamos a cabo el proceso de validación cruzada jackknife, que proporciona los resultados que se muestran en la siguiente figura:

Figura 3.2. Resultados jackknife lista PART - *Modelo 1*

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

Correctly Classified Instances          64      88.8889 %
Incorrectly Classified Instances         8      11.1111 %
Total Number of Instances               72

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.917     0.139      0.868     0.917     0.892    mala
  0.861     0.083      0.912     0.861     0.886    buena

=== Confusion Matrix ===

  a  b   <-- classified as
 33  3 |  a = mala
  5 31 |  b = buena
```

El elevado porcentaje estimado de acierto del 88.9% deja fuera de toda duda la capacidad predictiva del modelo. Como se puede observar, además del porcentaje global de acierto señalado, en la figura 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: $\text{TP Rate} = \frac{TP}{TP+FN}$, 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 $\text{FP Rate} = \frac{FP}{FP+TN}$, 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, $\text{Precision} = \frac{TP}{TP+FP}$, 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:

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

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.

Estudiemos ahora el significado de esta sencilla representación en forma de lista de la regularidad presente en los datos. La primera de las reglas hace referencia a los ratios R3 y R6. La condición relativa al ratio R3 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 la sección 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. En cuanto al ratio R6, recordemos que se trata del denominado "ratio de cobertura": R6 que 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. De acuerdo con la primera regla de la lista, cuando el activo circulante sea menor o igual que el 140% del pasivo circulante y además las provisiones técnicas no superen el 96% de los ingresos por primas, 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 cualquiera de estas dos condiciones no se verifica, lo que ocurrirá si el ratio de cobertura, o el de liquidez,

o ambos, son lo suficientemente elevados, no podemos concluir lo contrario – que la empresa sea sana –, sino que simplemente la regla deja de ser aplicable y pasamos a considerar la que viene a continuación. De acuerdo con ella, para que la empresa sea clasificada como "buena" se le exigirá un valor más alto en el ratio de cobertura, concretamente superior al 112%. Cuando esta segunda regla no fuese tampoco aplicable entrarían en juego los ratios R20 y R4. El ratio R20 es el denominado "ratio combinado", el cual, como se ha indicado en la sección precedente, ha sido establecido con carácter general dentro del análisis contable del sector del seguro como una medida de evaluación de la gestión global de la actividad aseguradora, y se define como la suma de otros dos ratios, el ratio de siniestralidad y el ratio de gastos de gestión. Cuando el ratio combinado tome un valor menor que la unidad, significará que la gestión de la empresa habrá sido eficiente. En cuanto al ratio R4, recordemos que se trata del denominado "ratio de garantía": R4 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". Como señala la tercera de las reglas de la lista de decisión, aunque el ratio de cobertura anterior no fuese elevado, lo cual tampoco implica que haya de ser necesariamente muy bajo por no haberse podido clasificar la empresa de acuerdo con las reglas precedentes, la eficiencia en la gestión empresarial acompañada de una señal consistente de distancia a la quiebra recomendarían considerar sana a la empresa.

Prestando atención al rango de valores que ha de tomar el ratio de distancia a la quiebra, esto es, debe ser mayor que 1.86 pero no superior a 2.32, se pone de manifiesto que si bien este ratio debe superar netamente la unidad, valores muy elevados tampoco proporcionarían una absoluta certidumbre a la hora de catalogar a la empresa como "buena", 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 podrían deberse a un escaso nivel de actividad o a una constitución incorrecta de provisiones técnicas que, en cualquier caso, no recomendarían clasificar la empresa como "buena", de ahí que la regla establezca un límite superior para este ratio.

La siguiente regla de la lista señala que, de no verificarse alguna o ninguna de las condiciones anteriores, las empresas cuya cuantía de provisiones técnicas supere el 37% de la de ingresos por primas y cuyo activo real sea mayor que el 146% de su pasivo exigible, serán "buenas". Si esta regla tampoco fuese aplicable, entonces aquellas empresas con R23 igual o inferior al –1%, lo que significa una cuantía de pérdidas del ejercicio (antes de impuestos) igual o superior al 1% de la cuantía de fondos propios, esto es, rentabilidad financiera negativa, serán "malas". Finalmente, si tampoco se verificase esta regla, la empresa será considerada fracasada.

Para el segundo año anterior a la quiebra, se ha optado por el valor fijado por defecto en el programa para el parámetro que establece el número de instancias que, como mínimo, debe cubrir cada una de las reglas de la lista, siendo éste igual a 2, y un valor del 5% para el parámetro que controla la intensidad de la poda de los árboles parciales construidos, de manera que cuanto menor sea el valor de dicho parámetro, más acusada será la poda de los árboles parciales que darán lugar a cada una de las reglas de la lista. Así, el modelo obtenido se presenta en la **Figura 3.3**.

Como se desprende de la lista, el número total de empresas clasificadas incorrectamente es de 12, lo que supone un acierto del 82.35% en la clasificación del conjunto de entrenamiento, porcentaje que se reduce ligeramente, en concreto al 80.88%, cuando es estimado mediante el método jackknife, como puede verse en la **Figura 3.4**. Este resultado muestra de nuevo que el modelo goza de buena capacidad predictiva.

Figura 3.3. Lista de Decisión Modelo 2

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

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

PART decision list
------------------

R14  <= 0.399419 AND
R3  > 1.002715 AND
R6  > 0.046776 AND
R6  > 0.906477: buena (12.0/1.0)

R3  <= 1.611955: mala (28.0/6.0)

R14  <= 0.633184: buena (18.0/3.0)

: mala (10.0/2.0)

Number of Rules  :      4
```

Figura 3.4. Resultados jackknife lista PART Modelo 2

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

Correctly Classified Instances          46      85.1852 %
Incorrectly Classified Instances         8      14.8148 %
Total Number of Instances               54

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.889     0.185      0.828     0.889      0.857    mala
  0.815     0.111      0.88      0.815      0.846    buena

=== Confusion Matrix ===

  a  b   <-- classified as
 24  3 |  a = mala
  5 22 |  b = buena
```

Esta segunda lista de decisión se compone de tan sólo cuatro reglas, la última de ellas la clase por defecto. La primera de las reglas consta de cuatro condiciones, aunque, como puede observarse, en realidad serían tres, ya que una de ellas está contenida en otra. Nos referimos a R6 > 0.05 y R6 > 0.91, siendo la primera redundante. Debido a la forma de operar del algoritmo PART, en ocasiones se observa este fenómeno de que aparecen condiciones redundantes en alguna de las reglas de la lista. Desde un punto de vista puramente estético, quizás convendría introducir algún cambio en la implementación del algoritmo de manera que una vez derivadas las reglas aquellas condiciones redundantes en las mismas fuesen eliminadas, y de este modo las listas de decisión presentasen un aspecto más compacto, aunque, obviamente, esto no afecta a los resultados en absoluto, y probablemente por ello los autores de PART no hayan prestado mucha atención a este asunto.

Tal y como se observa en la lista de decisión el algoritmo selecciona ahora como variables más discriminantes los ratios R14, R6 y R3. Por medio del ratio R14 se manifiesta la solvencia 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 la sección 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 como 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 periodo analizado.

De este modo, de acuerdo con la lista, cuando se cumpla que la siniestralidad de la empresa no supere el 40% de la suma de fondos propios y provisiones técnicas, el fondo de maniobra sea positivo y la cuantía de provisiones técnicas mayor que el 91% de la de ingresos por primas, la empresa se considerará "buena". Cuando no se verifique alguna o ninguna de las tres condiciones, las compañías cuyo ratio de liquidez no supere el 161% serán "malas", y si tampoco se cumple esta regla, y por tanto el ratio de liquidez es mayor del 161%, con un valor como máximo en el ratio de solvencia R14 del 63% la empresa será "buena". Finalmente, cuando ninguna de las tres reglas anteriores sea aplicable la empresa será "mala". Esto ocurrirá cuando la compañía, a pesar de no manifestar problemas de liquidez, y sea cual sea su ratio de cobertura, posea un valor alto para el índice de solvencia R14.

En las dos figuras que se muestran a continuación (**Figura 3.5** y **Figura 3.6**) se exponen, respectivamente, la lista de decisión derivada a partir de los datos del tercer año previo a la quiebra, la cual se ha obtenido exigiendo que cada regla de la lista cubra al menos 5 de las 54 empresas de que consta la muestra para este tercer año, así como los resultados proporcionados por el método jackknife. En la lista de decisión se puede observar que el total de errores en la clasificación es de 6, lo que supone un porcentaje de clasificaciones correctas sobre el conjunto de entrenamiento del 89%. Por su parte, en la figura que recoge los resultados de la validación cruzada jackknife, se observa que el porcentaje de acierto estimado es del 85%, lo que nos conduce a confiar en la bondad del modelo.

Figura 3.5. Lista de decisión. Modelo 3

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

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

PART decision list
------------------

R17  <= 0.447703: buena (9.0/1.0)

R12  <= 0.534979: mala (7.0)

R6  <= 0.857601 AND
R6  > 0.171021 AND
R12  <= 2.500576: mala (6.0)

R12  <= 2.262512 AND
R6  > 0.405206: buena (11.0)

R20  > 1.112924: mala (8.0/1.0)

R24  > 0.012519: buena (7.0/2.0)

: mala (6.0/2.0)

Number of Rules  :      7
```

Figura 3.6. Resultados Jacknife lista PART. Modelo 3

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

Correctly Classified Instances          55       80.8824 %
Incorrectly Classified Instances        13       19.1176 %
Total Number of Instances               68

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.824     0.206      0.8       0.824     0.812    mala
  0.794     0.176      0.818     0.794     0.806    buena

=== Confusion Matrix ===

  a  b   <-- classified as
 28  6 |  a = mala
  7 27 |  b = buena
```

Como refleja la primera de las reglas de la lista 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. Si así no fuere, es decir, cuando el ratio de siniestralidad sea mayor del 45%, la segunda regla indica que será "mala" la empresa cuya siniestralidad sea menor o igual que el 53% de la cuantía de fondos propios, y aunque desde el punto de vista de la solvencia de la empresa esta situación no sería criticable, podría ser indicativa de un nivel de actividad escaso en relación al volumen de fondos propios que de no corregirse conducirá a medio plazo a un deterioro patrimonial de la empresa y eventualmente a su liquidación.

Cuando no se verifique ninguna de las dos condiciones anteriores pasaremos a la tercera regla, representativa de la clase "mala". Observando el lugar que ocupa en la lista, una empresa que verificase esta regla tendría un valor para el ratio R12 del 250% como máximo, pero mayor del 53%, ya que si no se habría clasificado según la regla anterior. Entonces el ratio R12 podría no tomar valores muy extremos, lo que en principio sería positivo –puesto que tanto cuantías excesivas como escasas de fondos propios 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– , pero al no ser su ratio de cobertura superior al 86%, la empresa se consideraría "mala". Resulta chocante, sin embargo, que la regla establezca también la condición de que el ratio de cobertura deba superar el 17%, de manera que una empresa con un valor en este ratio de, por ejemplo, el 30%, verificando también la condición sobre el ratio R12 sería clasificada como "mala". En cambio, si el ratio de cobertura fuese del 10%, ¿ya no podría considerarse "mala"?. Esta condición un tanto absurda se explica por los datos en sí. Observando los valores del ratio en nuestra muestra para el tercer año, ocurre que muy pocas empresas toman un valor en su ratio de cobertura del 17% o inferior, pero la mayoría de ellas resultan ser "buenas", por ello el algoritmo selecciona este punto de corte con el fin de obtener una regla "limpia", sin fallos, que caracterice a las empresas fracasadas. Del análisis de las cuentas anuales del pequeño conjunto de empresas sanas con valores muy bajos en su ratio de cobertura se puede extraer la conclusión de que se trata de empresas de muy reducida dimensión que funcionan satisfactoriamente pero llevan su contabilidad de forma negligente e incompleta, presentando una evidente infradotación de provisiones.

En caso de no cumplirse alguna o ninguna de las condiciones de la regla anterior, la siguiente señala que si R12 no es superior al 226% (obviamente, será mayor del 53%) y el ratio de cobertura supera el 41%, la empresa será "buena". Los casos que verifiquen estas dos condiciones cumplirán también la tercera condición de la regla precedente ($R12 \leq 2.5$), así que tomarán en realidad un valor para el ratio de cobertura superior al 86%, por no haber verificado dicha regla precedente.

En definitiva, hasta aquí las reglas vienen a significar que cuando no se observe gran eficiencia en la gestión técnica y el ratio de solvencia R12 tome valores intermedios, será el valor del ratio de cobertura el que determine la supervivencia o no de la empresa a tres años vista, de manera que cuando éste sea elevado la empresa sobrevivirá y si no fracasará.

Cuando la cuarta regla tampoco sea aplicable, serán "malas" las empresas con un valor superior a 1.11 para el ratio combinado. En caso contrario, como indica la siguiente regla donde se evalúa el ratio R24, si el beneficio de la empresa es superior al 1% del total de su pasivo, ésta será "buena", y serán "malas" las empresas para las cuales no sea aplicable ninguna de las seis reglas anteriores.

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 número de reglas de las respectivas listas de decisión (incluida la clase por defecto) y el conjunto de ratios intervinientes en las mismas.

TABLA 3.1. Resultados de las listas de decisión PART.

Modelo	Ratios	Número de reglas	Clasificaciones correctas			
			Conjunto de entrenamiento		jackknife	
			Empresas "buenas"	Empresas "malas"	Empresas "buenas"	Empresas "malas"
1	R3,R6, R20, R4, R23	6	94.4%	100%		
			Total: 97.2%			
2	R14, R3, R6	4	76.5%	88.2%	79.4%	82.4%
			Total: 82.4%		Total: 80.9%	
3	R17, R12, R6, R20, R24	7	88.9%	88.9%	81.5%	88.9%
			Total: 88.9%		Total: 85.2%	

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.

3.2.1. 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 frente al Análisis Discriminante la ventaja 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 $\log\frac{p}{1-p}=\beta_0+\beta_1x_1+\ldots+\beta_px_p+\varepsilon$ 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 así 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 estandarizados resultan tener un valor absoluto superior a 100 son considerados extremos y convertidos en *missing*. 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 hace aconsejable 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 convierten, 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*. 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 nuestros datos, el valor más adecuado de *k* resulta ser 8 en los tres años.

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 $-2\log\left[L\left(\hat{\theta}\right)\right]+p\log n$ en donde *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 3.2. 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%	

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     0.6403   2.585  0.01063 *
R5           -1.1827     0.5213  -2.269  0.02329 *
R6           -2.6247     1.4438  -1.818  0.06907 .
R7            2.1665     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
```

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 listas de decisión PART, 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 PART 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 PART 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 convierte su utilización en especialmente atractiva 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 PART 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 periodo 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 PART 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.

Finalmente nos gustaría indicar que 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, la utilización de este método a modo de los *early warning systems* (sistemas de alerta temprana) empleados en Estados Unidos y en otros países de nuestro entorno facilitaría grandemente la labor de supervisión de las entidades aseguradoras.

Referencias Bibliográficas

Altman, E.I. (1968): "Financial ratios, discriminant analysis and the prediction of the corporate bankruptcy", The Journal of Finance, vol. 23, nº 4, pp. 589-609.

Altman, E.I.; Haldeman, R.G. y Narayanan, P. (1977): "ZETATM analysis: a new model to identify bankruptcy risk of corporations", Journal of Banking & Finance, vol. 1, nº 1, pp. 29-54.

Altman, E.I. y Loris, B. (1976): "A Financial Early Warning System for Over-the-Counter Broker-Dealers", The Journal of Finance, vol. 31, nº 4, pp. 1201-1217.

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, nº 1, pp. 113-122.

Berzal Galiano, F. (2002): ART: un método alternativo para la construcción de árboles de decisión. Tesis Doctoral, Universidad de Granada.

Best, A.M. Company (1991): Best's Insolvency Study, Property/Casualty Insurers 1969-1990. A.M. Best Company, Special Report, June 1991.

Breiman, L.; Friedman, J.H.; Olshen, R.A. Y Stone, C.J. (1984): Classification and Regression Trees. Wadsworth International Group, Belmont, California.

Deakin, E.B. (1972): "A Discriminant Analysis of Predictors of Business Failure", Journal of Accounting Research, vol.10, nº 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.

Efron, B. (1982): The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pennsylvania.

Fernández Palacios, J. Y Maestro, J.L. (1991): Manual de Contabilidad y Análisis Financiero de Seguros. Centro de Estudios del Seguro, Madrid.

Frank, E. Y Witten, I.H. (1998): "Generating Accurate

Rule Sets Without Global Optimization", en J. SHAVLIK (ed.): Proceedings of the Fifteenth International Conference on Machine Learning, Madison, Wisconsin. Morgan Kaufmann, San Francisco, pp. 144-151.

Gabás Trigo, F. (1990): Técnicas Actuales de Análisis Contable. Instituto de Contabilidad y Auditoría de Cuentas, Madrid.

Gabás Trigo, F. (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, nº 1, pp. 146-160.

Hampton, J. (1991): Financial management of insurance companies. PCG Publishing, New Jersey.

Hernández Orallo, J.; Ramírez Quintana, M.J. Y Ferri Ramírez, C. (2004): Introducción a la Minería de Datos. Pearson Prentice Hall, Madrid.

Jiménez Cardoso, S.M.; García-Ayuso Covarsí, M. Y Sierra Molina, G.J. (2000): Análisis financiero. Pirámide, Madrid.

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.

Laffarga Briones, J.; Martín Marín, J.L. Y Vázquez Cueto, M.J. (1987): "Predicción de la crisis bancaria española: la comparación entre el análisis logit y el análisis discriminante", Cuadernos de Investigación Contable, vol. 1, nº 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.

Lizarraga 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.

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, nº 49, Madrid.

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.

Mensah, Y.M. (1983): "The Differential Bankruptcy Predictive Ability of Specific Price Level Adjustments: Some Empirical Evidence", The Accounting Review, vol. 58, nº 2, pp. 228-246.

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, nº 78, pp. 203-233.

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, nº 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, Edinburgh, UK, pp. 168-201.

Quinlan, J.R. (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.

Quinlan, J.R. (1986): "Induction of decision trees", Machine Learning, vol. 1, nº 1, pp. 81-106.

Quinlan, J.R. (1993): C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, California.

R Development Core Team (2005): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, 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 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.

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, nº 14, Valladolid.

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, nº 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.

Sanchis Arellano, A.; 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, nº 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.

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, Pennsylvania, vol. 2, pp. 111-174.

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, nº 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.