Universidad Politécnica de Valencia

Facultad de Administración y Dirección de Empresas

Trabajo final de Máster en Dirección Financiera y Fiscal

Análisis del riesgo crediticio de las empresas españolas mediante el uso de “decision trees”

Curso 2013-2014

Autor:

Eduardo Alcácer Ribelles

Director:

Francisco Guijarro Martínez

Índice

Índice de figuras

Índice de tablas

1. Resumen

El objeto de este Trabajo Fin de Máster es la modelización de la situación de insolvencia empresarial a partir de la información económico-financiera de una muestra importante de empresas españolas, así como la propuesta de un modelo para la predicción de la insolvencia y el consecuente riesgo crediticio mediante el uso de una arquitectura de aprendizaje basada en árboles de decisión.

Dentro de las numerosas técnicas de aprendizaje automatizado se encuentran los árboles de decisión que son métodos estadísticos basados en sistemas expertos y cuyo objetivo es realizar una clasificación de los elementos para los que ha sido entrenado. En general son métodos muy rígidos y que necesitan de un entrenamiento supervisado para poder trabajar, además presentan el inconveniente de la necesitan de recalcularse desde cero en caso de querer clasificar nuevos elementos.

Para dar el carácter dinámico y adaptativo al sistema, el árbol de decisión se combina con otra técnica de aprendizaje automatizado conocida como agrupamiento o “clustering”. Esta técnica crea, a partir de un conjunto de elementos, diferentes grupos, agrupando elementos con características similares. En la arquitectura propuesta, dicha técnica se utiliza para generar grupos a partir de los elementos “desconocidos” para, con ellos, generar un nuevo árbol de decisión que incluya los nuevos objetos y permitir, de ese modo, una adaptación del sistema a elementos nuevos.

Los algoritmos de aprendizaje automatizado que se utilizarán en este trabajo son también similares a versiones modificadas de algoritmos ampliamente conocidos como son el ID3 o C4.5, para los árboles de decisión, y el “k-means” para el agrupamiento.

El trabajo se estructura de la siguiente manera.

En el primer punto se resumen las partes de este trabajo.

En el segundo punto se realiza una breve introducción donde se expondrá el contexto del trabajo, se estudian los antecedentes del mismo y se expone la motivación que ha llevado a la consecución del trabajo.

En el tercer punto se establecen los objetivos que vamos a perseguir en este trabajo.

En el cuarto punto se realiza una revisión del estado del arte, describiendo las técnicas de predicción de solvencia más conocidas y realizando una revisión del trabajo que otros investigadores y grupos han realizado.

En el quinto punto se describe el desarrollo del trabajo, desde la obtención de los resultados que servirán como punto de partida, hasta la explicación detallada de la arquitectura de aprendizaje propuesta así como los algoritmos desarrollados

En el sexto punto se exponen las pruebas realizadas y los resultados obtenidos durante la validación del sistema, analizando cada uno de los algoritmos de aprendizaje expuestos en el capítulo de desarrollo.

Por último se detallan las conclusiones extraídas del análisis de los resultados y la adecuada utilidad que este algoritmo proporciona en la predicción de la insolvencia empresarial.

2. Objetivos

Los objetivos principales planteados para este Trabajo Fin de Máster son los siguientes:

- Repasar los métodos y modelos utilizados en la medición del riesgo crediticio.
- Analizar el estado del arte en la predicción de la insolvencia empresarial.
- Implementar un algoritmo que permita predecir, de una manera autónoma la probabilidad de insolvencia de una empresa.
- Testear el funcionamiento del algoritmo de aprendizaje de árboles de decisión ID3 en un ejemplo sencillo para la predicción de la insolvencia de una compañía en función de varias variables financieras extraídas de los distintos estudios analizados, como las más determinantes para la estimación de la insolvencia empresarial en los años previos al concurso de acreedores o quiebra.
- Partiendo de los registros contables de la compañía en el año n-1 y el estado de la compañía en el estado n, se implementarán modelos predictivos para predecir la probabilidad de insolvencia futura que ayude al acreedor a la toma de decisión de dar el crédito y su riesgo asociado.
- A partir de lo anterior se desarrolla un sistema de rating empresarial para clasificar empresas, con la implementación de tres modelos :
    - Modelo de regresión logística
    - Modelo de máquinas de vector soporte (SVM)
    - Modelo de particionado recursivo
- Se destaca el modelo basado en árboles de decisión, mediante la técnica del particionado recursivo como el que da mejores resultados en la predicción de la insolvencia empresarial.

3. Introducción

3.1. Presentación del proyecto

El presente Trabajo Final de Máster (TFM) surge de la necesidad e importancia de tener una metodología adecuada para la predicción de la insolvencia empresarial de cara a los posibles riesgos crediticios que la situación de estas puede generar. Esta metodología ya ha sido estudiada por diversos autores y aplicada a empresas cotizadas, empresas del sector seguros y otras, pero que en este caso nos vamos a centrar en una técnica de clasificación y razonamiento basada en casos como son los árboles de decisión y concretamente aplicada a cualquier tipo de empresa, independientemente de sus características, aunque sí se realizará un análisis diferenciado según su tamaño, actividad, cifra de negocio y otros, que dará lugar a los diferentes "clusters" a partir de los que formaremos los distintos árboles de decisión.

La actividad de una entidad financiera es la toma de riesgos y cada una de sus operaciones contiene incertidumbre en las decisiones. En cada operación se expone a diferentes tipos de riesgo que deben ser identificados, medidos y controlados, como base sobre todo para fijar precios, que resulten de una ecuación favorable entre el riesgo asumido y la recompensa obtenida, medida como la rentabilidad neta.

En el contexto actual, el riesgo de crédito ha pasado a tener una importancia crucial y de la cual se pueden destacar las circunstancias siguientes:

1. La reforma de la regulación financiera internacional, iniciada, para las entidades bancarias, con la normativa establecida en el Nuevo Acuerdo de Capital de Basilea -Basilea II-, de reciente implantación, en la que la medición y gestión de los riesgos financieros a los que están expuestas dichas entidades tiene una trascendental importancia.

2. La fase de expansión que el ciclo crediticio ha experimentado en los últimos años y la brusca contracción que ha sufrido después y que se sigue soportando en la actualidad.

Entre los diversos factores causantes de dicha fase de expansión es posible identificar la incorrecta medición por las entidades bancarias del riesgo de crédito al que han quedado expuestas al invertir en activos financieros ilíquidos, lo que ha provocado efectos negativos tanto en la economía real –burbuja inmobiliaria- como en la financiera (excesivo endeudamiento de las empresas y las familias).

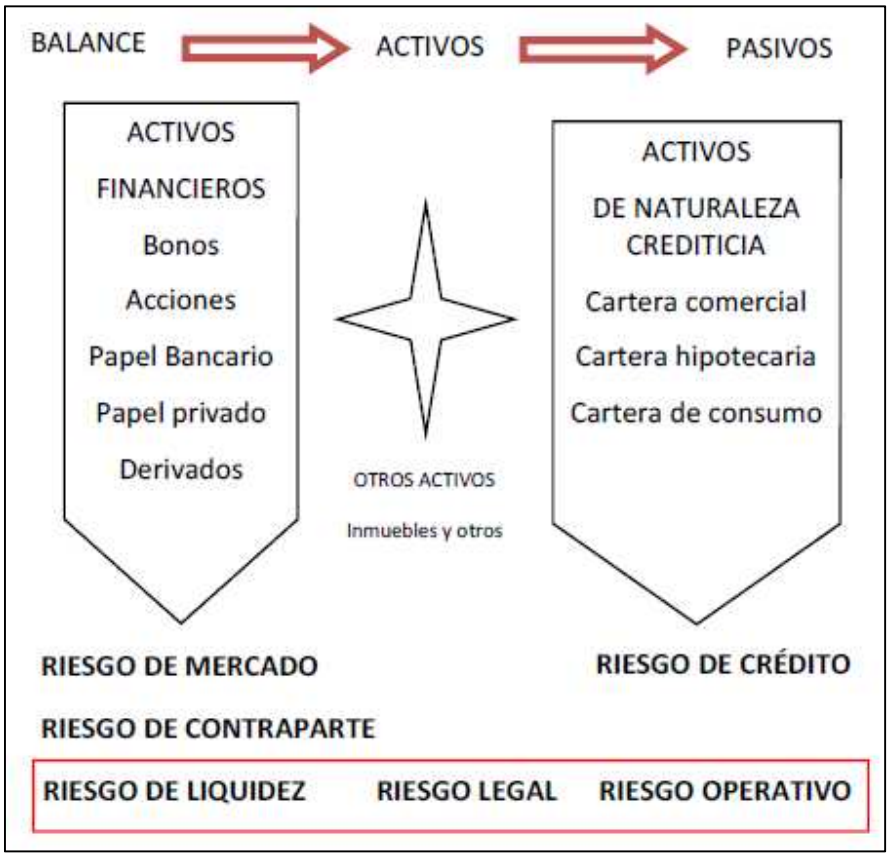
3. La actual crisis financiera internacional, iniciada precisamente en el sector bancario con la emisión de activos financieros ilíquidos de dudosa recuperación y su posterior difusión al resto del sistema financiero por medio de su titulización, la cual ha llevado a la quiebra a importantes entidades financieras y ha perjudicado gravemente la solvencia y viabilidad de otras muchas, poniendo en serio peligro la estabilidad del propio sistema financiero y en cuestión la labor que llevan a cabo las agencias de clasificación crediticia y los organismos de supervisión y control de la actividad financiera. Asimismo, dicha crisis financiera está teniendo graves consecuencias tanto en la economía real –destrucción de tejido empresarial e incremento del desempleo- como en la financiera (reducción del crédito a las empresas e incremento de la deuda pública).

En los últimos años, las entidades financieras han pasado de gestionar el riesgo mediante métodos empíricos, a hacerlo con metodologías que se apoyan en métodos estadísticos.

La forma de valorar técnicamente el riesgo es mediante la cuantificación del grado de variación de los resultados reales frente a los estimados. A mayor volatilidad, mayor incerteza para el accionista o acreedor.

Los distintos tipos de riesgo a los que se enfrenta una entidad financiera y que son frecuentemente analizados son entre otros:

Ilustración 1: Riesgo en las operaciones activas de balance



- Riesgo de liquidez: Incumplimiento de un compromiso financiero con un tercero por no poder deshacer una posición determinada al precio de mercado o no poder cumplir con compromisos de pago en la fecha de cancelación.
- Riesgo legal: es la contingencia de pérdida derivada de situaciones de orden legal.
- Riesgo operativo: posibilidad de pérdida como resultado de deficiencias a causa de fallas en los sistemas de información, procesos, errores humanos, control gerencial, etc.
- Riesgo de mercado: Es el riesgo de que una o más variables relevantes para la entidad financiera, cuyo valor depende de los mercados financieros, evolucionen de forma adversa a las expectativas de ésta, provocándole pérdidas. Entre los más importantes tenemos:

  - Riesgo de tipo de interés: Es el riesgo de que una variación en los tipos de interés provoque pérdidas en las operaciones financieras que realiza la entidad financiera.
  - Riesgo de tipo de cambio: Es el riesgo de que una variación en la relación entre dos monedas ocasione pérdidas en las operaciones financieras que la entidad financiera lleva a cabo. El riesgo de tipo de cambio afecta a las entidades financieras que operan en mercados financieros internacionales empleando fuentes de financiación y realizando inversiones denominadas en monedas distintas a su moneda nacional.
- Riesgo de contraparte: es la posibilidad de incumplimiento de las obligaciones contractuales entre la entidad financiera y el sector financiero
- Riesgo de crédito: posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos como consecuencia de que sus deudores fallen en el cumplimiento de los términos contractuales acordados.

Cada uno de los riesgos, tiene su método y forma de medición.

En este TFM nos centraremos en el riesgo de crédito y la utilización de “decisión trees” para predecir la probabilidad de incumplimiento o probabilidad de insolvencia de la compañía a financiar o a invertir.

La necesidad de medir el riesgo y promover que las instituciones financieras hagan una correcta evaluación de ellos ha sido un esfuerzo de todos los bancos a nivel mundial. El comité de supervisión bancaria de Basilea, ha sido precursor de la reglamentación de la medición integral de riesgos y el adecuado aprovisionamiento de capitales, para sobrellevar los posibles riesgos incurridos y evitar la quiebra de las instituciones financieras. Se aspira a que todos los bancos internacionalmente apliquen las recomendaciones consignadas en el acuerdo de capitales de Basilea que definen el capital mínimo que deben tener las entidades financieras para operar, relacionando el riesgo de los activos con el nivel y calidad del patrimonio, además, determina el capital mínimo necesario para que un banco haga frente a posibles quebrantos debido a los riesgos que maneja.

Para el cálculo de la probabilidad de incumplimiento o de insolvencia, existen diferentes metodologías estadísticas con las que es posible predecir la probabilidad de insolvencia o fallo en un periodo dado. Entre las metodologías están el análisis discriminante, regresión logística, análisis probit, redes neuronales, matrices de transición y árboles de decisión, que son las más comúnmente usadas en el sector financiero.

En este TFM se presenta el uso de árboles de decisión como herramienta para el cálculo de probabilidades de insolvencia, por cuanto representa un método efectivo para la estimación, al igual que la mayoría de los métodos alternativos, pero ofrece la ventaja fundamental al ser un método de fácil entendimiento para personas que no cuentan con conocimientos avanzados de estadística. El modelo permite realizar una clasificación de clientes considerados como buenos y de máxima solvencia (probabilidades bajas de incumplimiento) y hasta clientes considerados como malos con altas probabilidades de insolvencia.

3.2. Antecedentes del TFM

El fracaso empresarial es uno de los problemas que ha venido enfrentando la economía lo largo del tiempo y sobre el cual aún no se ha llegado a elaborar una teoría positiva que permita la anticipación del mismo. Por ello, y desde una óptica puramente empírica, se han elaborado diversos sistemas de indicadores de alerta de crisis financieras.

Debido a la gran cantidad de causas que pueden llevar a la discontinuidad de la actividad de una empresa, las cuales no siempre están presentes en todo fracaso empresarial, y a la falta de conocimiento de cómo relacionar éstas con la posterior insolvencia de la empresa, aún no se ha llegado a establecer una teoría general del fracaso empresarial, que permita alertar con antecedencia sobre el mismo, de modo a evitar las indeseables consecuencias que conlleva.

En este sentido, la anticipación del fracaso empresarial viene siendo objeto de numerosos estudios a fin de permitir que todos aquellos relacionados económicamente con la empresa cuenten con herramientas que permitan detectar la insolvencia con antecedencia.

Onusic y Casa Nova (2006) destacan que los modelos de predicción de insolvencia son de grande auxilio en la evaluación del riesgo, siendo utilizados como una herramienta importante en el análisis de desempeño de las empresas y apoyo a las decisiones de crédito.

Ontiveros y Valero (1987, p. 25) denominan a los modelos de predicción de crisis empresariales “modelos microeconómicos”, dado que éstos parten de la base de que los principales problemas del fracaso empresarial se localizan en la misma empresa, de forma que se tratará, en definitiva, de determinar los rasgos más significativos que pueden observarse en las empresas en crisis, por contraposición con las empresas con éxito, rasgos que sirvan para detectar, e incluso predecir, la situación de fracaso.

La naturaleza del deudor es de gran importancia en la medición del riesgo de crédito porque de ella dependen las variables fundamentales en la medición del riesgo de crédito.

Dicha naturaleza influye en la definición de la variable aleatoria “estado en el que se encuentra el deudor”, en los métodos y modelos, ya sean empíricos o teóricos, que pueden emplearse en la determinación de su distribución de probabilidad, en las variables explicativas que pueden utilizarse en estos modelos y en la muestra o población que puede emplearse en la determinación de sus parámetros.

Asimismo, la naturaleza del deudor también influye en la variable aleatoria “pérdida en caso de impago”, ya que de ésta depende la capacidad que tiene la entidad financiera para exigirle al deudor el cumplimiento de sus obligaciones en caso de impago, salvaguardando de esa forma sus intereses.

El riesgo de crédito aquí estudiado es el de las empresas privadas, por lo que a continuación se exponen las características de este tipo de empresas que son más relevantes para la medición de dicho riesgo.

El sector privado está formado por las empresas cuyo capital es de propiedad privada.

Las empresas pueden clasificarse en función de diversos criterios. Uno de los criterios más importantes es el tamaño, en función del cual se distingue entre empresas grandes, medianas y pequeñas.

El Comité de Supervisión Bancaria de Basilea emplea un criterio basado en los tres elementos siguientes:

1. El tamaño de la empresa, medido por el importe anual de sus ventas.
2. La exposición de la entidad financiera al riesgo de crédito de la empresa.
3. El método utilizado por la entidad financiera en la medición el riesgo de crédito de la empresa.

En función de los criterios expuestos más arriba, este organismo establece la siguiente clasificación de empresas:

1. Las grandes empresas, que son aquéllas cuyo importe anual de ventas es igual o superior a cincuenta millones de euros.

2. Las pequeñas y medianas empresas, que son aquéllas cuyo importe anual de ventas es inferior a cincuenta millones de euros.

Dicho organismo distingue, a su vez, entre dos tipos de pequeñas y medianas empresas: uno compuesto por aquellas empresas con un importe anual de ventas inferior a cincuenta millones de euros, y otro tipo formado por aquellas empresas en las que, además, la entidad financiera presenta una exposición al riesgo de crédito inferior a un millón de euros y emplea una metodología de medición de este riesgo similar a la que se emplea en las familias.

Esta clasificación es relevante por las diferencias que presentan las grandes y las pequeñas y medianas empresas en la medición del riesgo de crédito.

3.2.1. Métodos y modelos de medición del riesgo crediticio y probabilidad de insolvencia de las empresas.

Los principales métodos y modelos de medición del riesgo de crédito de los activos financieros pueden clasificarse en función de su naturaleza en los tres siguientes:

1. Los métodos basados en el factor humano.
2. Los modelos estadísticos.
3. Los modelos de aprendizaje de máquinas.

3.2.1.1. Métodos basados en el factor humano

Los métodos basados únicamente en el factor humano son los empleados tradicionalmente por las entidades bancarias.

En dichos métodos, un empleado o grupo de empleados de la entidad financiera analiza un conjunto de variables explicativas de las variables fundamentales en la medición del riesgo de crédito y las ponderan en función de su criterio personal con la finalidad de medir el riesgo de crédito del deudor, del activo financiero o de ambos.

Dichos empleados han demostrado, a lo largo de su carrera profesional, poseer las habilidades y la experiencia necesarias para medir el riesgo de crédito de determinados activos financieros y/o deudores, por lo que son considerados expertos en la materia.

Las variables explicativas que emplean dichos expertos suelen agruparse en cinco categorías son las siguientes:

1. Capital, que agrupa aquellas variables explicativas, fundamentalmente de tipo cuantitativo, que determinan la situación patrimonial del deudor.

2. Capacidad, que comprende el conjunto de variables explicativas, fundamentalmente de tipo cuantitativo, que determinan los recursos financieros que el deudor es capaz de generar periódicamente, así como su dispersión.

3. Carácter, que agrupa aquellas variables explicativas, principalmente de tipo cualitativo, que determinan la disposición del deudor a cumplir con sus obligaciones.

4. Colateral, que engloba el conjunto de variables explicativas relacionadas con la variable aleatoria “pérdida en caso de impago”.

5. Ciclo, que comprende el conjunto de variables explicativas relacionadas con la influencia del estado del ciclo económico en las variables fundamentales en la medición de riesgo de crédito.

También están los métodos de medición del riesgo de crédito basados en el factor humano y en normas, que son aquéllos en los que la entidad financiera determina el proceso que deben llevar a cabo sus empleados para medir el riesgo de crédito de los activos financieros y de los deudores, estableciendo un conjunto de principios, normas, formularios e instrumentos que éstos deben utilizar con tal fin.

Dichos métodos son más objetivos que los basados únicamente en el factor humano. No obstante, ambos son intensivos en este factor, por lo que suelen presentar los mismos inconvenientes, los cuales son, principalmente, los cinco siguientes:

1. La medición del riesgo de crédito es heterogénea, ya que depende de factores personales del empleado que son ajenos al activo financiero y al deudor, tales como,

por ejemplo, el estado anímico y los posibles perjuicios del empleado, o los factores ambientales del lugar en el que se realice la medición del riesgo de crédito.

2. El departamento de medición del riesgo de crédito no controla directamente los métodos basados en el factor humano, lo que dificulta el control y la supervisión, así como la realización de modificaciones.

3. Estos métodos no pueden informatizarse, por lo que el coste, el tiempo y el trabajo que requieren, así como la posibilidad de comisión de errores, son mayores que en el resto de los métodos y modelos de medición del riesgo de crédito.

La capacidad de medición del riesgo de crédito de los empleados de la entidad financiera es limitada y se adquiere por medio de un proceso de aprendizaje que habitualmente suele ser largo, lo que exige de la entidad un doble proceso de planificación: el del volumen de créditos y el del personal necesario para medir su riesgo.

4. El proceso de aprendizaje del personal cualificado fomenta un mayor grado de concentración en la cartera, ya que este personal se especializa en la medición del riesgo de crédito de uno o más tipos de activos financieros o deudores.

3.2.1.2. Modelos estadísticos

Los modelos estadísticos pueden clasificarse en univariantes y multivariantes, en función del número de variables explicativas que se emplean simultáneamente en la medición del riesgo de crédito.

3.2.1.2.1. Modelos estadísticos univariantes

Los primeros modelos estadísticos utilizados en la medición del riesgo de crédito de las empresas fueron los modelos univariantes.

Se trata de modelos caracterizados porque en cada momento se emplea una única variable independiente, lo cual no significa que el riesgo de crédito se mida empleando una sola variable, sino que, en el caso de que se utilicen dos o más variables, la medición se realizaría de forma secuencial, es decir, midiendo el riesgo por medio de las diferentes variables independientes en primer lugar y, a continuación, agregando los

resultados obtenidos en cada una de ellas por medio de un determinado método, el cual proporciona una medida global del riesgo de crédito.

La principal ventaja que presentan estos modelos es su sencillez, mientras que sus principales inconvenientes son los siguientes:

1. El riesgo de crédito es un fenómeno complejo y multidimensional, que depende de numerosas variables explicativas, por lo que no puede representarse con exactitud empleando una sola variable.

2. La medición del riesgo de crédito por medio de modelos estadísticos univariantes y dos o más variables independientes requiere que ésta sea secuencial, lo que supone un conflicto cuando dos o más variables proporcionan resultados contrapuestos.

No obstante, los modelos estadísticos univariantes se utilizan ampliamente en la selección de las variables explicativas, ya que la capacidad explicativa univariante es un requisito para la multivariante.

Dentro de esta rama de modelos destacan:

- Análisis de tendencias: se limitan a analizar la evolución de los ratios financieros utilizados durante un determinado periodo, con la finalidad de determinar si existen diferencias entre las empresas solventes e insolventes. La principal conclusión que se obtiene es que los ratios financieros presentan un valor en las empresas solventes distinto del que presentan en las empresas insolventes, diferencia que es mayor cuanto más avanzado se encuentra el proceso de fracaso empresarial o más próximo se encuentra el suceso impago.
- Análisis de ratios financieros: El análisis de ratios financieros es uno de los modelos más utilizados en la medición del riesgo de crédito de las empresas y consiste en comparar el valor que un determinado ratio presenta en la empresa con el que tiene en una muestra o población de empresas, de forma que si es mayor –o menor, dependiendo del ratio financiero de que se trate- la empresa se clasifica como solvente y en caso contrario como insolvente. La principal ventaja del análisis de ratios financieros es su simplicidad, mientras que sus principales

inconvenientes son que no permite determinar los ratios financieros que son más adecuados para la medición del riesgo de crédito ni la validez de la clasificación o predicción obtenida con su uso.

El primer trabajo en el que se determinan cuáles son los ratios financieros más adecuados para la medición del riesgo de crédito es el de Beaver (1966), en el que se determinan estos ratios empleando el test de clasificación dicotómica y el análisis de los ratios de probabilidad. Esto lo veremos junto con un estudio actualizado en el apartado del estado del arte.

- Test de clasificación dicotómica. es un método que permite clasificar las observaciones de una muestra o población en dos grupos. Las principales ventajas del test de clasificación dicotómica son su sencillez y que el análisis de los errores de clasificación por medio de tablas de contingencia permite determinar cuáles son los ratios financieros más adecuados para la medición del riesgo de crédito. En cuanto a los inconvenientes, los principales son La dificultad que conlleva la determinación del punto de corte óptimo, y que carece de métodos y medidas de validación.
- Análisis de los ratios de probabilidad: este análisis proporciona una medida que relaciona el suceso impago con el valor de un determinado ratio financiero, la cual no es otra que la variable aleatoria “estado en el que se encuentra el deudor, condicionado al ratio financiero”. La probabilidad asociada a esta variable aleatoria se obtiene por medio del teorema de la probabilidad condicionada. Todas estas probabilidades pueden obtenerse representando la muestra de empresas en dos histogramas: uno para las empresas en las que se produce el suceso pago y otro para aquéllas en las que se produce el impago.

  En el eje de abscisas de estos histogramas se representa los valores de un determinado ratio financiero y en el de ordenadas las frecuencias relativas de aquellas empresas cuyo ratio financiero presentan estos valores.

  La probabilidad de que la variable aleatoria “estado en el que se encuentra el deudor” presente el valor pago –o el impago- puede estimarse observando el histograma de las empresas en las que se ha producido el suceso pago (o el impago). Asimismo, la probabilidad de que la variable aleatoria “ratio financiero,

condicionado al estado en el que se encuentra el deudor" presente un determinado valor puede estimarse observando el valor de este ratio financiero en el histograma de las empresas en las que se ha producido el suceso pago (o impago).

3.2.1.2.2. Modelos estadísticos multivariantes.

Los modelos estadísticos multivariantes se caracterizan porque emplean de forma simultánea dos o más variables explicativas en la medición del riesgo de crédito. Los principales modelos estadísticos multivariantes empleados en la medición del riesgo de crédito son el análisis discriminante, los modelos de regresión y los de supervivencia.

- Análisis discriminante: es un modelo estadístico multivariante que permite clasificar los elementos de una muestra o población en dos o más grupos mediante la utilización de una serie de reglas de clasificación. Los primeros trabajos en los que se aplica el análisis discriminante múltiple a la medición del riesgo de crédito de las empresas utilizan una regla de clasificación lineal distinguiéndose dos grupos, uno en el que esta variable aleatoria presenta el valor pago y otro en el que presenta el valor impago.

  Las deficiencias específicas del análisis discriminante múltiple están relacionadas con la incorrecta aplicación de este modelo a la medición del riesgo de crédito, siendo los principales errores que se comenten los siguientes:

  - La definición errónea de los grupos a clasificar. El análisis discriminante múltiple asume la hipótesis de que los grupos a clasificar son discretos, identificables y mutuamente excluyentes entre sí, hipótesis que se incumplen con frecuencia en la medición del riesgo de crédito.
  - El incumplimiento de la hipótesis de normalidad multivariante. En el análisis discriminante se asume la hipótesis de que la función de distribución de las variables independientes condicionada a que los elementos pertenezcan a uno de los grupos establecidos a priori es elíptica con carácter general, y normal multivariante en particular. Sin embargo, esta hipótesis no suele cumplirse realmente en la medición del riesgo de crédito.

- El incumplimiento de la hipótesis de igualdad de las matrices de varianzas-covarianzas.
- La dificultad que presenta la interpretación económica del signo y del valor absoluto de los coeficientes de las variables independientes, puesto que no son únicos
- La dificultad que supone la reducción dimensional, cuya finalidad es maximizar la distancia entre las medias aritméticas de los grupos, minimizando el número de variables independientes que componen la función discriminante.
- La dificultad que conlleva la elección del punto de corte óptimo.

3.2.1.2.3. Modelos de regresión

Entre estos tenemos los siguientes:

- Modelo de probabilidad lineal: es un modelo de regresión lineal multivariante en el que el fenómeno aleatorio se representa por medio de una variable dependiente dicotómica que presenta el valor uno cuando acaece el fenómeno y el cero en caso contrario. La estimación de los parámetros del modelo de probabilidad lineal puede llevarse a cabo por medio del método de los mínimos cuadrados ordinarios, para lo cual es necesario asumir una serie de hipótesis sobre el comportamiento del error aleatorio, entre las que destacan las siguientes: 1. Su esperanza matemática es cero. 2. Su varianza es constante (hipótesis de homocedasticidad). 3. No existe auto correlación.

  Los principales inconvenientes del modelo de probabilidad lineal son los siguientes: 1. La relación lineal entre las variables dependientes e independientes es una hipótesis que con frecuencia no se cumple en la medición del riesgo de crédito. 2. La existencia de heterocedasticidad supone que las estimaciones que proporciona el método de los mínimos cuadrados ordinarios son ineficientes.

  Asimismo, la utilización del método de los mínimos cuadrados generalizados incrementa los cálculos y no garantiza la obtención de unas estimaciones eficientes, sobre todo en muestras de tamaño reducido. 3. Las variables independientes tienen un ratio de compensación constante, característica que no

es deseable en la medición del riesgo de crédito. El ratio de compensación indica qué variación debe experimentar una variable independiente para compensar el efecto que la variación de otra variable independiente produce en la dependiente, de forma que el valor de esta última permanezca constante. 4. La puntuación crediticia puede presentar valores extraños –inferiores a cero o superiores a uno- en algunos elementos de la muestra o población.

- Modelo logístico y probabilístico: Los modelos de regresión logística y probabilística son modelos de regresión multivariante en los que el fenómeno aleatorio se representa por medio de una variable dependiente observable que depende de una variable dependiente inobservable. La variable dependiente observable presenta el valor uno cuando acaece el fenómeno objeto de estudio y el cero en caso contrario. Las hipótesis asumidas sobre la distribución de probabilidad del error aleatorio permiten obtener distintos modelos de regresión multivariante. Así, si se asume la hipótesis de que esta distribución de probabilidad es logística se obtiene el modelo de regresión logística, también denominado logit. Por el contrario, si se asume la hipótesis de que la distribución de probabilidad del error aleatorio es una normal univariante se obtiene el modelo de regresión probabilística, también denominado probit o normit. Los modelos de regresión logística y probabilística presentan diversas diferencias y similitudes. Entre las similitudes destaca que las distribuciones de probabilidad de ambos modelos son simétricas, monótonas crecientes, casi lineales en el centro y presentan dos asíntotas: una por la izquierda en el valor cero y otra en la derecha en el valor uno. Las principales diferencias entre ambos modelos son que la función de distribución del modelo de regresión probabilística queda por encima del modelo de regresión logística en las colas, que la varianza del modelo de regresión probabilística puede ser constante o variable, mientras que la del modelo de regresión logística siempre es constante, y que la obtención de la distribución de probabilidad en el modelo de regresión probabilística requiere integrar, mientras que la modelo de regresión logística no.

  Las principales ventajas del modelo de regresión logística son las siguientes:

1. Las probabilidades que proporciona este modelo pertenecen al intervalo [0,1] y, por consiguiente, pueden interpretarse como tales sin necesidad de llevar a cabo operaciones adicionales. 2. La relación entre las variables dependiente e independientes es no lineal, lo cual es una ventaja en un fenómeno como el riesgo de crédito en el que dicha relación es con frecuencia no proporcional. 3. En comparación con el análisis discriminante, presenta las dos ventajas siguientes: a) El signo de los coeficientes indican el sentido del efecto que una variación de las variables independientes produce en la dependiente. b) La estimación de los parámetros es más robusta, aunque también es cierto que es menos eficiente. 4. En relación con el modelo de regresión probabilística, la determinación de la distribución de probabilidad no requiere integrar, lo que facilita su obtención.

3.2.1.2.4. Modelos de aprendizaje de máquinas

Los modelos de aprendizaje de máquinas pueden clasificarse en función del algoritmo que emplean en la medición del riesgo de crédito en modelos de inteligencia artificial, de redes neuronales artificiales y de árboles de decisión.

- Modelos de inteligencia artificial. Los modelos de inteligencia artificial deducen, principalmente por inducción, el procedimiento que llevan a cabo los mejores empleados de la entidad financiera en la medición del riesgo de crédito, reproduciéndolo mediante un programa informático. El principal elemento de estos modelos es una base de datos, denominada de conocimiento, que está compuesta por un conjunto de reglas, de algoritmos y de variables dependientes e independientes. Las reglas, denominadas de producción, son del tipo "si … entonces …", estableciendo relaciones de causa y efecto entre las variables dependientes e independientes, de la forma siguiente: si "el grado de apalancamiento financiero del deudor es mayor o igual a x" entonces "el crédito resulta impagado".

  Estos algoritmos son desarrollados conjuntamente por estos empleados y los expertos en inteligencia artificial, de forma que los primeros describen los procedimientos que emplean en la determinación del valor de las variables

fundamentales en la medición del riesgo de crédito y los segundos los transcriben en reglas de producción.

- Modelos de redes neuronales artificiales: Las redes neuronales artificiales son un modelo no paramétrico que trata de reproducir la inteligencia humana imitando el funcionamiento del cerebro humano y de sus unidades básicas: las neuronas. De forma similar al cerebro humano, las redes neuronales artificiales están formadas por una o varias redes de elementos conectados entre sí imitando el funcionamiento de las neuronas biológicas, denominados neuronas artificiales, procesadores elementales o nodos, los cuales son dispositivos de cálculo simple que reciben unas entradas, realizan una serie de operaciones y proporcionan un resultado.
- Modelos de árboles de decisión: son unos modelos no paramétricos y secuenciales, cuya finalidad es descomponer un proceso de toma de decisiones complejo en un conjunto de decisiones más simples. Para ello, se divide la muestra o población en dos o más grupos en función de un determinado criterio, con el objeto de que sean más homogéneos que los originales. Los grupos obtenidos se dividen, a su vez, en otros dos o más subgrupos en función de otro criterio, con el fin de que sean más homogéneos que los primeros. El proceso se repite el número de veces que sea necesario para alcanzar la finalidad expuesta. La finalidad del diseño de un árbol de decisión es obtener una estructura que sea lo más simple posible, fácilmente actualizable, que clasifique la muestra de entrenamiento de forma exacta y que sea generalizable, es decir, que cuando se aplique a muestras distintas de la de entrenamiento proporcione resultados aceptables. Atendiendo al número de grupos que se obtienen de las sucesivas divisiones, los árboles de decisión se clasifican en: 1. Binarios, que son aquéllos en los que se obtienen dos grupos. 2. No binarios, que son aquéllos en los que se obtienen más de dos grupos. Cualquier árbol de decisión no binario puede obtenerse por medio de una combinación de árboles binarios. De los diferentes tipos de formación de los árboles de decisión, el que más se utiliza es el de divisiones iterativas (recursive partitioning algorithm).

Las principales ventajas de los modelos de árboles de decisión son las siguientes: 1. Son modelos no paramétricos. 2. Pueden representarse gráficamente. 3. Permiten la división del espacio muestral en espacios que no tienen por qué ser conexos entre sí. 4. Son simples y fáciles de comprender, controlar, mantener y usar. 5. Admiten el uso de información incompleta, ruidosa y con errores sistemáticos. 6. Permiten la utilización de variables independientes cuantitativas, ya sean continuas o discretas, cualitativas o una combinación de ambas. Asimismo, una variable independiente puede emplearse en más de una división.

Los principales inconvenientes de los modelos árboles de decisión son los siguientes: 1. Asumen la hipótesis de que las categorías predeterminadas son discretas, identificables y que no se superponen entre sí. 2. Realizan una división ortogonal del espacio muestral, lo que supone un inconveniente cuando presenta una estructura diagonal. 3. No permiten determinar la importancia que cada una de las variables independientes tiene en la clasificación de los elementos en las categorías predeterminadas. 4. Las reglas de división son secuenciales, con el inconveniente que ello supone cuando se utilizan reglas de parada basadas en el podado del árbol de decisión, puesto que la eliminación de un nodo padre conlleva la eliminación de todos los nodos hijos que quedan por debajo. 5. Las reglas de parada que limitan directamente el crecimiento del árbol de decisión proporcionan árboles de decisión óptimos que se ajustan con exceso a la muestra de entrenamiento, presentando una capacidad de generalización limitada y menoscabando los resultados de las otras muestras. 6. La determinación del árbol de decisión óptimo resulta difícil, dependiendo de numerosos elementos, muchos de los cuales son difíciles de determinar, destacando las probabilidades marginales, el coste asociado a la mayor complejidad del árbol o el coste de clasificar en una categoría los elementos que pertenecen a otra.

3.3. Motivación del Trabajo Fin de Máster

Tradicionalmente, para abordar el problema de la detección precoz de la insolvencia empresarial, se han venido utilizando métodos estadísticos que emplean ratios financieros como variables explicativas. Sin embargo, aunque la eficacia de dichos métodos ha sido sobradamente probada, presentan algunos problemas que dificultan su aplicación en el ámbito empresarial, ya que, generalmente, se trata de modelos basados en una serie de hipótesis sobre las variables explicativas que en muchos casos no se cumplen y, además, dada su complejidad, puede resultar difícil extraer conclusiones de sus resultados para un usuario poco familiarizado con la técnica.

El presente trabajo describe una investigación de carácter empírico consistente en la aplicación del algoritmo de inducción de reglas y árboles de decisión a partir de un conjunto de ratios financieros de una muestra de empresas españolas, con el objeto de comprobar su utilidad para la predicción de insolvencias en este sector. También se comparan los resultados alcanzados con los que se obtienen aplicando otras metodologías. Estas técnicas, procedentes del campo de la Inteligencia Artificial, no presentan los problemas mencionados.

El objetivo esencial para este trabajo es el logro de un mejor aprovechamiento de la información financiero-contable suministrada por las entidades sometidas a supervisión que permitan 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 impida hacer frente a los compromisos adquiridos.

4. Estado del Arte

En el siguiente apartado se realizará una revisión de los estudios de los diferentes métodos existentes para la predicción precoz de la insolvencia empresarial.

Como antecedentes más relevantes tenemos estudio de Gabás (1990), que estableció como criterio de solvencia un parámetro en función de la cotización ser superior o igual al 50% o inferior al 50% para establecer las muestras de empresas sanas y fracasadas. Las técnicas empleadas por Gabás (1990) fueron el análisis discriminante, análisis logit y algoritmo de particiones sucesivas, alcanzando resultados con la muestra de estimación, con las tres técnicas utilizadas, superiores al 94%, siendo el análisis logit el que produjo el porcentaje de acierto mayor (98%).

A nivel internacional, los precursores de las investigaciones estadísticas sobre modelos de predicción de fracaso fueron Beaver (1966), a través de técnicas univariantes y Altman (1968), con técnicas multivariantes. Desde entonces, se han realizado una gran cantidad de estudios sobre el tema, estando entre los más importantes, desarrollados para empresas que incluyan a las del sector industrial como en este estudio, los de Altman (1968) y (1993), Deakin (1972), Edmister (1972), Blum (1974), Taffler (1974) descrito en Taffler (1982), Libby (1975), Taffler y Tisshaw (1977), Altman *et al* (1977), Ohlson (1980), Gentry *et al* (1985) y Casey y Bartczak (1984) y (1985).

El criterio mayoritariamente adoptado en dichos estudios para definir el fracaso ha sido el de quiebra legal o suspensión de pagos, con excepción de Edmister (1972) que consideró fracasadas a las empresas que no devolvieron un préstamo. En cuanto al tamaño de las empresas, los estudios de Altman (1968) y (1993), Taffler (1974), Taffler y Tisshaw (1977) y Ohlson (1980) trabajaron con empresas cotizadas, siendo que los restantes trabajos se basaron en empresas de tamaño pequeño y medio.

Las técnicas multivariantes utilizadas fueron en casi todos ellos el análisis discriminante, con excepción de Ohlson (1980) que fue uno de los pioneros en utilizar el análisis logit. Gentry *et al* (1985) utilizó, además del análisis discriminante, el análisis logit y el análisis probit y Casey y Bartczak (1984) y (1985) también aplicaron el análisis

logit, además del discriminante. Los resultados alcanzados en dichos estudios con la muestra de estimación y con los datos del año anterior al fracaso se encuentran entre 87% y 99% y aquellos que validaron los modelos obtuvieron resultados para el año anterior al fracaso entre 72% y 95%, siendo, por tanto, como es de esperar más bajos que los de estimación.

4.1. Información contable en la predicción de insolvencia: estudio inferencial univariante aplicado a empresas españolas.

En este estudio se realizan estimaciones que permiten establecer relaciones significativas entre determinados ratios financieros y la insolvencia empresarial. La muestra envuelve empresas españolas que cotizan en Bolsa no pertenecientes al sector financiero y de seguros. Los resultados indican que varios ratios contables consiguen clasificar las empresas en solventes e insolventes con un grado de acierto del 95% para el año anterior al fracaso. Los indicadores de Rentabilidad del Activo, Margen de Beneficio del Resultado Ordinario y Cobertura de Gastos Financieros fueron los que resultaron con mayor poder discriminante. Se concluye, por lo tanto, que modelos univariantes pueden presentar expresiva capacidad predictiva del riesgo de insolvencia, además de contar con mayor practicidad en su proceso de estimación e implementación, en contraposición a los modelos multivariantes.

4.1.1. Estudios previos sobre modelos univariantes de predicción de fracaso empresarial.

El estudio más clásico, basado en ratios contables, es el de Beaver (1966) que, a pesar de su antigüedad, conserva todavía su vigencia metodológica. El objetivo del estudio de Beaver (1966) es la predicción de la insolvencia empresarial a través de ratios, concluyendo que el ratio Cash flow a Deuda total es el de mayor valor predictivo. Las técnicas utilizadas fueron la comparación de las medias para obtener un perfil de las empresas fracasadas y no fracasadas en los años anteriores al fracaso y el test de clasificación dicotómica, que se basa en un único ratio y representa, a diferencia del anterior, un test predictivo.

Se introduce a continuación un cuadro resumen con los estudios sobre modelos univariantes de insolvencia de empresas y la técnica aplicada.

Tabla 1: Investigaciones empíricas más relevantes sobre modelos de predicción univariantes de insolvencia de empresas industriales

Autor	Comp. Muestra (F y S)	Sectores	Emparej.	Período	Ratios seleccio-nados	Técnica Estad.	Rtdos. M. Est.
Beaver (1966)	79 y 79 (de Estados Unidos)	38 sectores industriales diferentes	Tamaño de activo y tipo de industria	1954-1964	Cash Flow/ Deuda Total (CF/DT) e Benef. Neto/ Activo Total (BN/AT)	Clasif.Dicotómica	n-1: CF/DT: 90% BN/AT: 88% n-2: CF/DT: 82% BN/AT: 85%
Deakin (1972)	32 y 32 de estimación (de Estados Unidos)	Industriales	Tamaño de activo, tipo de industria y año	1964-1970	Aplica los 14 ratios de Beaver con mejores rtdos.	Clasif. Dicotómica (y también el Análisis Discriminante)	CF/DT es el que presenta mayor poder de clasif. n-1: 80% y n-2: 84%
Wilcox (1973)	52 y 52 (de Estados Unidos)	Industriales	Tamaño de activo, tipo de industria y año	1955-1971	Mejores ratios de Beaver	Cadenas de Markov	Mejores rtdos. que Beaver n-1: 94% n-2: 90%
Lev (1973)	Igual muestra Beaver	Igual muestra Beaver	Igual muestra Beaver	Igual muestra Beaver	Cash Flow/ Deuda Total (CF/DT)	Anál. p/descompos. basa-da en la compar.de dos masas patrimoniales	Ratio CF/DT como el de mayor poder discriminante
Moyer (1977)	Igual muestra Beaver	Igual muestra Beaver	Igual muestra Beaver	Igual muestra Beaver	Cash Flow/ Deuda Total (CF/DT)	Anál. p/descompos. basa-da en la compar.de dos masas patrimoniales	Ratio CF/DT como el de mayor poder discriminante

4.1.2. Metodología

La presente investigación se caracteriza como cuantitativa y descriptiva.

Para alcanzar este objetivo se parte de un análisis descriptivo sobre las características económico-financieras de las empresas que fracasan, frente a las que no lo hacen. Seguidamente, y con el fin de identificar un perfil económico-financiero de ambos grupos de empresas a partir de las cuentas anuales, se aplican técnicas estadísticas que permitan identificar qué variables diferencian o clasifican mejor a ambos grupos de empresas y permitan avalar la separación previa de la muestra en los dos grupos de empresas.

La metodología utilizada para realizar el estudio descriptivo del comportamiento de ambos grupos de empresas es a través de instrumentos de análisis exploratorio, tales

como: medidas de tendencia central (media y mediana); medidas de dispersión (desviación típica, coeficiente de variación, mínimos y máximos); y medidas de distribución (asimetría y curtosis o apuntamiento).

Seguidamente, se profundiza en la investigación empírica a través del proceso inferencial univariante para identificar y definir las variables que individualmente consideradas permitan explicar y predecir la insolvencia empresarial.

Beaver (1966) fue el precursor en la aplicación de la técnica univariante para la estimación de modelos de predicción de insolvencia y lo realizó por medio del Análisis Dicotómico, el cual consiste en ordenar todas las observaciones de la muestra por cada ratio analizado y encontrar un punto de corte que separe a las empresas insolventes de las solventes de forma tal que se minimice el número de observaciones incorrectas. Posteriormente, autores como Deakin (1972) y Elam (1975) también aplicaron la técnica de Análisis Dicotómico empleada por Beaver (1966) como parte de sus investigaciones, a los cuales les siguieron otros autores, destacándose en España los estudios de Lizarraga (1997), García *et al* (1998) y Somoza (2000).

Primeramente, se identifican los ratios con diferentes medias entre los grupos a través de la técnica Análisis de la Varianza de un factor (ANOVA), a fin de conocer los posibles ratios con poder discriminante. Seguidamente se realiza el Análisis de Perfiles de Beaver (1966) en base a los resultados del ANOVA a fin de definir un perfil económico-financiero que muestre las diferencias, estadísticamente significativas, entre los dos colectivos estudiados, a fin de corroborar las hipótesis que surgieron del análisis descriptivo.

El Análisis de Perfiles de Beaver (1966), es una representación de cada uno de los ratios para cada estado de la empresa, fracasada o sana, y es aplicado en cada año anterior al fracaso. Éste permite analizar la diferencia entre ambos tipos de empresas teniendo en cuenta sólo la media, no así la dispersión, no proporcionando una idea sobre la magnitud o cuantificación de dichas diferencias.

Finalmente, se seleccionan los mejores modelos predictivos univariantes a través del Análisis Dicotómico de Beaver (1966). La prueba de clasificación dicotómica de Beaver

(1966) es un test predictivo, cuya finalidad última es seleccionar que ratio permite una mejor discriminación entre ambos grupos de empresas. Esta prueba se aplica a cada año anterior a la insolvencia. El punto de corte se determina mediante un proceso de prueba y error, por el cual se van fijando diferentes valores y se va tanteando cuál de ellos produce menores errores de clasificación, de tal forma que si:

$$Ratio_i \leq punto\ de\ corte = clasificación:\ fracasada$$

$$Ratio_i > punto\ de\ corte = clasificación:\ sana$$

Por tanto, para cada ratio se determina el punto de corte que separa a las empresas fracasadas de las sanas, observándose para cada uno el porcentaje de acierto y de error en la clasificación de ambos grupos de empresas. Los Tipos de Error ante una clasificación dicotómica pueden ser:

Error Tipo I: clasificar una empresa fallida como sana

Error Tipo II: clasificar una empresa sana como fallida

Las consecuencias de incurrir en un tipo u otro de error son claramente distintas, así por ejemplo en el caso de concederse un préstamo o decidir invertir en una empresa, el coste de error Tipo I es mucho mayor que el Tipo II, puesto que el Tipo II es un coste de oportunidad asociado a la no elección de dicha empresa; en cuanto el Tipo I involucra la pérdida de parte o totalidad del capital invertido.

Por ello, dependiendo del caso, resulta aconsejable intentar minimizar el porcentaje de error Tipo I o Tipo II, en vez del porcentaje de error total. Por lo que en el análisis de la solvencia debería intentarse reducir el error Tipo I, ya que éste provoca consecuencias más serias que el Tipo II.

Para la selección de los mejores modelos univariantes, se parte de los ratios con mejor porcentaje de acierto en la clasificación de los grupos de empresas resultante del Análisis Dicotómico, analizándose luego las correlaciones, a través del coeficiente de correlación de Pearson, a fin de evaluar la dependencia entre cada una de las variables consideradas.

4.1.3. Descripción de la muestra y variables consideradas

Las empresas objeto del estudio son empresas españolas cotizadas en Bolsa no pertenecientes al sector financiero ni de seguros. El criterio de fracaso utilizado en este estudio fue el correspondiente a la delimitación legal, por el cual se consideran fracasadas aquellas empresas en las que se ha admitido la solicitud de Suspensión de Pagos o Quiebra.

De acuerdo con Gabás (1997, p. 15) la variedad de situaciones por las que puede transitar una empresa insolvente dificulta fuertemente dar el concepto de fracaso empresarial, lo que obliga a los investigadores de la insolvencia o del fracaso empresarial a definir su concepto propio de forma explícita, por lo que se utilizan variadas definiciones en función de los objetivos o en razón de la disponibilidad de datos.

Las empresas fracasadas fueron identificadas a partir de la lectura de los hechos relevantes comunicados a la Comisión Nacional del Mercado de Valores en el período 1992 a 2001, de cada una de las 691 empresas no financieras ni de seguros supervisadas por la Comisión que comprendían la muestra total. De esta forma se identificaron 30 empresas fracasadas. Por ser necesario para la elaboración de modelos de predicción de insolvencia, se procedió a realizar, al igual que autores como Beaver (1966), Wilcox (1973), Altman (1968), Deakin (1972), Edminster (1972), Blum (1974), Taffl er (1974) descrito en Taffl er (1982), Taffler y Tisshaw (1977), Altman *et al* (1977), Gentry *et al* (1985), Lizarraga (1997), García (Coord.) (1997), Gallego *et al* (1997), Ferrando y Blanco (1998), entre otros, el emparejamiento de cada empresa fracasada con otra que no haya fracasado del mismo sector, tamaño, en función del activo total, y año, conformando así una muestra total de 60 empresas. Finalmente, se procede a la obtención de las cuentas anuales de la muestra de empresas fracasadas y de la muestra de empresas emparejadas de los dos años previos al fracaso.

Las variables a ser consideradas corresponden a dos grupos, uno integrado por magnitudes estructurales de la muestra y otro compuesto por una serie de ratios financieros.

En conjunto totalizaron 79 variables, siendo 6 de ellas magnitudes estructurales y 73 ratios contables. Las variables estructurales estudiadas fueron: Activo Total, Ventas, Activo Fijo, Fondos Propios, Resultado Neto y Personal.

En el Tabla 2 se exponen los ratios económico-financieros y en la Tabla 3 se presenta la descripción de las nomenclaturas utilizadas en los ratios.

La agrupación de los ratios, mostrada en la Tabla 2, se ha realizado, bajo un criterio simplificador, en cuatro grupos principales, por lo que los ratios de estructura del activo se encuentran dentro del grupo de ratios de Liquidez, los de estructura del pasivo dentro de los ratios de Endeudamiento y los ratios de generación de recursos han sido agrupados junto con los ratios de Rentabilidad.

Tabla 2: Ratios económico-financieros del estudio Información contable en la Predicción de la insolvencia. Douglas, Taboada y Lima.

Nº	Nombre de la variable	Descripción	Agrup.	Nº	Nombre de la variable	Descripción	Agrup.
1	RATIO01	AC/PC	Liq.	38	RATIO38	(GP + DA)/VA	Rent.
2	RATIO02	AD/PC	Liq.	39	RATIO39	GP/V	Rent.
3	RATIO03	(T + IFT)/PC	Liq.	40	RATIO40	RF/V	Rent.
4	RATIO04	T/PC	Liq.	41	RATIO41	V/AT	Rot.
5	RATIO05	AD/AC	Liq.	42	RATIO42	V/AF	Rot
6	RATIO06	T/AC	Liq.	43	RATIO43	V/FP	Rot
7	RATIO07	INTÉRV.S/CRÉD.	Liq.	44	RATIO44	V/DT	Rot
8	RATIO08	(AC - PC - E)/APR	Liq.	45	RATIO45	V/PC	Rot
9	RATIO09	(T + IFT)/CC	Liq.	46	RATIO46	AC/V	Rot
10	RATIO10	CC/(V + OING + RF)	Liq.	47	RATIO47	CC/V	Rot
11	RATIO11	AC/AT	Liq.	48	RATIO48	AD/V	Rot
12	RATIO12	AD/AT	Liq.	49	RATIO49	E/V	Rot
13	RATIO13	(T + IFT)/AT	Liq.	50	RATIO50	D/V	Rot
14	RATIO14	T/AT	Liq.	51	RATIO51	T/V	Rot
15	RATIO15	CC/AT	Liq.	52	RATIO52	PC/DT	End.
16	RATIO16	RN/AT	Rent.	53	RATIO53	FP/AT	End.
17	RATIO17	RAT/AT	Rent.	54	RATIO54	FP/DT	End.
18	RATIO18	RAIT/AT	Rent.	55	RATIO55	PC/FP	End.
19	RATIO19	RA/AT	Rent.	56	RATIO56	ELP/AT	End.
20	RATIO20	RAIT y Ext./AT	Rent.	57	RATIO57	PC/AT	End.
21	RATIO21	RE/AT	Rent.	58	RATIO58	DT/AT	End.
22	RATIO22	VA/AT	Rent.	59	RATIO59	RP/PT	End.
23	RATIO23	CFT/AT	Rent.	60	RATIO60	RP/PC	End.
24	RATIO24	RN/FP	Rent.	61	RATIO61	FP/AF	End.
25	RATIO25	RAT/FP	Rent.	62	RATIO62	RP/AF	End.
26	RATIO26	RN/DT	Rent.	63	RATIO63	AR/DT	End.
27	RATIO27	RAIT/DT	Rent.	64	RATIO64	DA/IM	End.
28	RATIO28	CFT/DT	Rent.	65	RATIO65	DA/(AF – IF + GDVE)	End.
29	RATIO29	RN/PC	Rent.	66	RATIO66	BNOD/AT	End.
30	RATIO30	RAT/PC	Rent.	67	RATIO67	GF/VA	End.
31	RATIO31	CFT/PC	Rent.	68	RATIO68	GF/V	End.
32	RATIO32	RN/V	Rent.	69	RATIO69	RAIT/GF	End.
33	RATIO33	RAT/V	Rent.	70	RATIO70	RE/GF	End.
34	RATIO34	RA/V	Rent.	71	RATIO71	VM/DT	Otros
35	RATIO35	CFT/V	Rent.	72	RATIO72	CFO/AT	Otros
36	RATIO36	CFT/RE	Rent.	73	RATIO73	% Crec.VA	Rent.
37	RATIO37	GF/DT	Rent.				

Tabla 3: Nomenclatura de los Ratios económico-financieros del estudio Información contable en la Predicción de la insolvencia. Douglas, Taboada y Lima.

Nomenclatura	Descripción
AC	Activo Circulante
AD	Activos Defensivos = AC – E
AF	Activo Fijo = Inmovilizado
APROV	Aprovisionamientos
AR	Activo Real = AT - Activos Ficticios
	Activos Ficticios = GDVE + Acc.p/Desemb.No Ex. + Acc.p/Desemb.Ex. + Acc.Propias CP
AT	Activo Total
BNOD	Beneficios No Distribuidos = Reservas = [Reservas (no R.Revaloriz. ni Prima Emis.) + Rtdos.Ej.Ant.]
CC	Capital Circulante = AC - PC = Fondo de Maniobra
CFO	Cash Flow Operativo calculado a partir del CFT = CFT - Variac. AC Explot. + Variac. PC Explot.
	AC Explotación = E + D + Ajustes Periodificación
	PC Explotación = PC - Deudas con Entidades de Crédito - Emisiones Obligaciones y Otros
CFT	Cash Flow Tradicional = RN + DA + Provis. LP = Recursos Generados
D	Deudores
DA	Dotación p/Amortizaciones de Inmovilizado
DT	Deuda Total = Pasivo Exigible = PT - FP = Fondos Ajenos = ELP + PC
E	Existencias
ELP	Exigible Largo Plazo = Acreed.a LP + Prov.Riesgos/Gtos.LP + Ing.a Dist.Vs.E.
FP	Fondos Propios = Patrimonio Neto
GDVS	Gastos a Distribuir en Varios Ejercicios
GF	Gastos Financieros
IF	Inmovilizado Financiero
IFT	Inversiones Financieras Temporales
IM	Inmovilizado Material
INTÉRV.S/CRÉD.	Intérvalo sin Crédito = (AC - E - PC) / (Gtos.Explot. - DA - Provis.de Tráfico)
OING	Otros Ingresos
PC	Pasivo Circulante + Prov.Riesgos/Gtos.CP
PT	Pasivo Total = FP + ELP + PC
RA	Resultado de la Actividad = Rtdo. de las Actividades Ordinarias
RAIT	Rtdo. antes de Int. e Imp. = RAT + GF
RAIT y Ex.	Rtdo. antes de Int., Imp.y Extraordinarios = Rtdo. Actividad + GF
RAT	Rtdo. antes de Impuestos
RE	Rtdo. de la Explotación = Ingresos Exp. – Gastos Exp.
RF	Rtdo. Financiero
RN	Rtdo. Neto = Rtdo. después de Impuestos = Rtdo. del Ejercicio
RP	Recursos Permanentes = Pasivo Fijo = FP + ELP
T	Tesorería = Disponible
V	Ventas = Importe Neto de la Cifra de Negocios
VA	Valor Añadido = Ingresos Exp. - Consumos Exp. - Otros Gtos. Exp. = RBE = Rtdo. Bruto de Exp.
	Consumos Exp. = Reducc. Exist. Prod. + APROV + GP
VM	Valor de Mercado. Cap. Social = Valor Bursátil de la Empresa = Valor de la Acción a FCE por cant. Acciones

4.1.4. Resultados y comparación con otros estudios

Del análisis descriptivo de las variables estructurales que caracterizan a la muestra de empresas fracasadas y sanas surgen rasgos diferenciales entre ambos grupos de empresas, considerando los dos años observados.

Con relación a los ratios contables, se desprende del análisis descriptivo la existencia de diferencias entre las medias de los ratios de los dos colectivos, observándose que las empresas fracasadas presentan valores menores en los ratios de liquidez, rentabilidad y rotación y superiores en los de endeudamiento, lo que permite realizar una primera aproximación sobre los rasgos económico-financieros que caracterizan en forma diferente a ambos grupos de empresas, que serán objeto de contraste estadístico a través de técnicas univariantes.

Para conocer si existen diferencias significativas en las medias de los ratios de las empresas fracasadas y las sanas, se realizó en primer lugar la prueba paramétrica de análisis de la varianza, conocido como ANOVA de un factor.

El resultado arrojó un total de 40 ratios, presentados en el Tabla 4, sobre los 73 estudiados, distribuidos en las cuatro categorías en que se han clasificado: liquidez, rentabilidad, rotación y endeudamiento, que rechazan la hipótesis nula de igualdad de medias, siendo éstos justamente los que demostraron no rechazar la normalidad según la prueba no paramétrica de Kolmogorov-Smirnov para una muestra, test que resultó menos estricto que los paramétricos.

Tabla 4: ANOVA de los Ratios económico-financieros del estudio Información contable en la Predicción de la insolvencia. Douglas, Taboada y Lima.

Nº	Nombre de la Variable	Nº	Nombre de la Variable	Nº	Nombre de la Variable	Nº	Nombre de la Variable
1	R01	11	R18	21	R30	31	R57
2	R02	12	R19	22	R31	32	R58
3	R04	13	R20	23	R37	33	R59
4	R06	14	R21	24	R41	34	R61
5	R12	15	R22	25	R42	35	R62
6	R13	16	R23	26	R44	36	R63
7	R14	17	R26	27	R45	37	R66
8	R15	18	R27	28	R53	38	R69
9	R16	19	R28	29	R54	39	R70
10	R17	20	R29	30	R56	40	R71

Luego se aplicó el Análisis de Perfiles a los ratios que mostraron diferencias significativas en las medias resultantes de la prueba ANOVA.

Del análisis de perfiles de cada ratio muestra en relación a las cuatro categorías en que se ha clasificado los ratios:

A) LIQUIDEZ: Todos los ratios de liquidez (R01, R02, R04, R06, R12,R13, R14 y R15) muestran que las empresas fracasadas presentan menor liquidez que las sanas a lo largo del período estudiado, empeorando levemente la misma cuando se acerca la fecha de fracaso. Por su parte la evolución de los ratios de liquidez en las empresas sanas se manifiesta en términos generales bastante constante, si bien puede observarse que los ratios R01, R02, R04, R06 y R15 exhiben una tendencia constante o ascendente y los ratios R12, R13 y R14 presentan una leve tendencia a la baja.

B) RENTABILIDAD: Al igual como ocurre con los ratios de liquidez, la totalidad de los ratios de rentabilidad (R16, R17, R18, R19, R29, R21, R22, R23, R26, R27, 28, R29, R30, R31 y R37) revelan menor rentabilidad en las empresas fracasadas y un caída bastante acentuada de ésta al acercarse la fecha de fracaso, en cuanto, las empresas sanas se mantienen constantes en los dos años o con una muy leve caída. Merece comentar el ratio R37, de gastos financieros a deuda total, en donde se percibe nítidamente el camino opuesto que toman ambos grupos de empresas respecto al mismo, el cual disminuye en las sanas al pasar del segundo año anterior al fracaso al primer año previo y aumenta en las fracasadas.

C) ROTACIÓN: Los ratios de rotación (R41, R42, R44 y R45) exhiben menor rotación en las empresas fracasadas, con una muy leve tendencia a la baja en ambos grupos de empresas.

D) ENDEUDAMIENTO: Los ratios de endeudamiento (R53, R54, R56, R57, R58, R59, R61, R62, R63, R68, R69, R70 y R71) denotan en todos ellos mayores niveles de endeudamiento en las empresas fracasadas que en las sanas.

Además, en la mayoría de los ratios, las empresas fracasadas presentan un aumento del mismo al acercarse a la fecha de fracaso, en cuanto, las empresas sanas se mantienen más o menos constantes.

Concluyendo sobre el perfil económico-financiero que caracteriza a los dos colectivos estudiados, se observa que las empresas fracasadas presentan valores bastante menores en los ratios de liquidez, rentabilidad y rotación y superiores en los de endeudamiento.

Sin embargo, dado que el test de comparación de medias y los gráficos de análisis de perfiles sólo revelan diferencias en las medias de los valores de los ratios pero no tienen en cuenta la dispersión que puede existir en cada uno de ellos, se aplicó el Análisis Dicotómico a fin de poder extraer conclusiones sobre la capacidad discriminante de los ratios, de modo a responder la pregunta de investigación sobre cuáles son los índices contables que pueden alertar sobre el riesgo de insolvencia y cuáles son los valores críticos de esos índices en los años previos a la discontinuidad.

De los resultados del análisis dicotómico, se observa que 22 ratios alcanzan un porcentaje de acierto global superior al 85% en el año anterior al fracaso, por lo que se procedió a realizar el análisis dicotómico de dichos ratios para el segundo año anterior al fracaso.

Del análisis de los resultados alcanzados, se desprende que los ratios R19 y R70 son los que proporcionan mejores resultados a nivel univariante. Ambos arrojan un porcentaje de acierto global de 95% para el año inmediato anterior al fracaso y de 83% para el segundo año anterior. Luego se encuentran los ratios R27 y R34, con un porcentaje de acierto global de 93% para el año anterior al fracaso y de 83% para el segundo año anterior.

También existen otros ratios con buen poder predictivo, los cuales si bien arrojan un porcentaje menor de acierto global que los anteriormente presentados, continúan superiores al porcentaje de 90% de acierto global para el año anterior al fracaso y al 80% para el segundo año anterior. Ellos son, por orden de mejores resultados alcanzados, los ratios R69, R22, R26, R30, R20, R18, R23, R16 y R28.

Por lo expuesto, y a fin de elegir el mejor ratio que discrimine a las empresas fracasadas y sanas, se seleccionó el ratio R19 de resultado de la actividad a activo total como la mejor variable univariante, ya que arroja un porcentaje de error Tipo I menor que el R70 y, como fue comentado previamente, se considera este error más grave que el Tipo II.

Buscando elegir cuál o cuáles de los otros ratios podrían ser considerados también dentro de los mejores indicadores, se analizó las correlaciones existentes entre ellos, a fi n de seleccionar aquellos que posean los mayores porcentajes de acierto y no estén altamente correlacionados, presentándose en la Tabla 5 los coeficientes de correlaciones de Pearson obtenidos. En base al análisis de las correlaciones entre los ratios con mejores poderes discriminantes, se eligieron como mejores modelos univariantes estimados los expuestos en la Tabla 5, los cuales presentan baja correlación entre ellos.

Tabla 5: Ratios económico-financieros univariantes mejores del estudio Información contable en la Predicción de la insolvencia. Douglas, Taboada y Lima.

Ratio	Descrip.	n-1					n-2				
		Punto de corte	% Acierto global	% Error total	% Error Tipo I	% Error Tipo II	Punto de corte	% Acierto global	% Error total	% Error Tipo I	% Error Tipo II
R19	RA/AT	0,004337	95,00	5,00	3,33	6,67	0,0316548	83,33	16,67	10,00	23,33
R70	RE/GF	1,015264	95,00	5,00	3,33	6,67	0,9929656	83,33	16,67	23,33	10,00
R34	RA/V	0,00294	93,33	6,67	6,67	6,67	0,0034288	83,33	16,67	26,67	6,67

Entre los ratios seleccionados como mejores modelos univariantes se encuentra en primer lugar el ratio de rentabilidad económica R19, compuesto por el Resultado de las Actividades Ordinarias a Activo Total, luego continúa el ratio R70 de cobertura de la deuda formulado por el Resultado de la Explotación a Gastos Financieros, y como último seleccionado, el ratio R34 de margen de beneficios compuesto por el Resultado de las Actividades Ordinarias a Ventas.

Los ratios seleccionados reflejan que los indicadores que captan con altísima capacidad de acierto los síntomas claves del progresivo fracaso de una empresa son ratios de rentabilidad y de endeudamiento. Así, el ratio R19 de Rentabilidad del Activo

evidencia la importancia para la continuidad de la empresa de mantener una relación adecuada entre la rentabilidad ordinaria y el activo total.

Ya el segundo ratio elegido, R70 de Resultado de las Actividades Ordinarias antes del Resultado Financiero (Resultado de la Explotación) a Gastos Financieros demuestra el rol importante que juegan los malos planteamientos financieros en la insolvencia posterior de la empresa. Finalmente, el tercer ratio escogido, R34 de Margen de Beneficio del Resultado Ordinario, denota la necesidad de conquistar al menos un mínimo de margen de beneficio para mantener la salud financiera de la empresa.

4.2. Evaluación del riesgo de insolvencia mediante el Análisis discriminante y el Análisis Logit.

En este estudio Laura Edith Taboada Pinheiro y Juliano Lima Pinheiro (2008) analizan las características económico-financieras de una muestra de empresas españolas industriales y con acciones en Bolsa, compuesta por empresas que tuvieron dificultades financieras y por empresas consideradas solventes, con el propósito de construir modelos de predicción de crisis empresariales. Las técnicas estadísticas que emplean son el Análisis Discriminante y el Análisis Logit. Ambos métodos proporcionaron altos porcentajes de acierto en la clasificación de las empresas, siendo éstos en torno del 95% para el año anterior al fracaso y del 80% en el segundo año anterior. El modelo Logit presentó resultados en la estimación levemente superiores a los del Análisis Discriminante, si bien en las pruebas de validación del modelo el Análisis Discriminante se mostró mejor, manteniendo los mismos porcentajes de acierto de la estimación. Los ratios seleccionados por los modelos como mejores variables explicativas del fracaso fueron, principalmente: los de rentabilidad económica y los de generación de recursos. De esta forma, el estudio permitió trazar un perfil característico de las empresas españolas que atravesaron problemas de insolvencia.

4.2.1 Descripción de la muestra

Las empresas objeto del estudio son empresas españolas de gran dimensión sujetas al control de la Comisión Nacional del Mercado de Valores (CNMV), no pertenecientes al sector financiero ni de seguros.

El primer paso para conformar la muestra fue definir el criterio de fracaso. En esta investigación se ha elegido, por su objetividad, el correspondiente a la delimitación legal, por el cual se consideraron fracasadas aquellas empresas en las que se ha admitido la solicitud de suspensión de pagos o quiebra. Para la identificación de las empresas fracasadas, se analizó cada uno de los hechos relevantes comunicados a la

CNMV en el período 1992 a 2001, de cada una de las 691 empresas no financieras ni de seguros supervisadas por la Comisión que comprendían la muestra total.

De esta forma se identificaron 30 empresas fracasadas. Por ser necesario para la elaboración de modelos de predicción de insolvencia, se realizó el emparejamiento de cada empresa fracasada con otra que no haya fracasado del mismo sector, tamaño, en función del activo total, y año, conformando así una muestra total de 60 empresas. Finalmente, se procedió a la obtención de las cuentas anuales de la muestra de empresas fracasadas y de la muestra de empresas emparejadas de los dos años previos al fracaso a fin de construir la base de datos en Excel.

4.2.2. Variables consideradas

La elección del conjunto posible de variables explicativas del fracaso empresarial se realizó en función de la revisión de varias investigaciones empíricas relevantes sobre modelos de predicción de fracaso como también de la revisión de la literatura de análisis contable, de modo de incluir todos aquellos ratios más frecuentemente utilizados.

De esta forma fue elaborada de la misma forma que en el estudio anterior, la lista de 73 ratios que hicieron parte del análisis, presentados en la tabla 1, agrupados, bajo un criterio simplificador, en cuatro grupos principales: Liquidez (L), Endeudamiento (E), Rentabilidad (R) y Rotación (Ro). Sus nomenclaturas se encuentran expuestas en los cuadros siguientes

Tabla 6: Ratios utilizados en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

Ratio	Descripción	Agr.	Ratio	Descripción	Agr.	Ratio	Descripción	Agr.
R01	AC/PC	L	R26	RN/DT	R	R51	T/V	Ro
R02	AD/PC	L	R27	RAIT/DT	R	R52	PC/DT	E
R03	(T+IFT)/PC	L	R28	CFT/DT	R	R53	FP/AT	E
R04	T/PC	L	R29	RN/PC	R	R54	FP/DT	E
R05	AD/AC	L	R30	RAT/PC	R	R55	PC/FP	E
R06	T/AC	L	R31	CFT/PC	R	R56	ELP/AT	E
R07	INTÉRV.S/CRÉD.	L	R32	RN/V	R	R57	PC/AT	E
R08	(AC-PC-E)/APR	L	R33	RAT/V	R	R58	DT/AT	E
R09	(T+IFT)/CC	L	R34	RA/V	R	R59	RP/PT	E
R10	CC/(V+OING+RF)	L	R35	CFT/V	R	R60	RP/PC	E
R11	AC/AT	L	R36	CFT/RE	R	R61	FP/AF	E
R12	AD/AT	L	R37	GF/DT	R	R62	RP/AF	E
R13	(T+IFT)/AT	L	R38	(GP+DA)/VA	R	R63	AR/DT	E
R14	T/AT	L	R39	GP/V	R	R64	DA/IM	E
R15	CC/AT	L	R40	RF/V	R	R65	DA/(AF–IF+GDVE)	E
R16	RN/AT	R	R41	V/AT	Ro	R66	BNOD/AT	E
R17	RAT/AT	R	R42	V/AF	Ro	R67	GF/VA	E
R18	RAIT/AT	R	R43	V/FP	Ro	R68	GF/V	E
R19	RA/AT	R	R44	V/DT	Ro	R69	RAIT/GF	E
R20	RAIT y Ext./AT	R	R45	V/PC	Ro	R70	RE/GF	E
R21	RE/AT	R	R46	AC/V	Ro	R71	VM/DT	Ot
R22	VA/AT	R	R47	CC/V	Ro	R72	CFO/AT	Ot
R23	CFT/AT	R	R48	AD/V	Ro	R73	%Crec.VA	R
R24	RN/FP	R	R49	E/V	Ro			
R25	RAT/FP	R	R50	D/V	Ro			

Tabla 7: Nomenclatura en los Ratios utilizados en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

Nomenclatura	Descripción
AC	Activo Circulante
AD	Activos Defensivos = AC - E
AF	Activo Fijo = Inmovilizado
APROV	Aprovisionamientos
AR	Activo Real = AT - Activos Ficticios
	Activos Ficticios = GDVE + Acc.p/Desemb.No Ex. + Acc.p/Des.Ex. +
	Acc.Propias CP
AT	Activo Total
BNOD	Ben. No distrib.. = Reservas = [Reservas (no R.Revaloriz. ni Prima Emis.) + Rtdos.Ej.Ant.]
CC	Capital Circulante = AC - PC = Fondo de Maniobra
CFO	Cash Flow Oper. calc. a partir del CFT = CFT - Variac. AC Explot. + Variac. PC Explot.
	AC Explotación = E + D + Ajustes Periodificación
	PC Explotación = PC - Deudas con Entidades de Crédito - Emisiones Obligaciones y Otros
CFT	Cash Flow Tradicional = RN + DA + Provis. LP = Recursos Generados
D	Deudores
DA	Dotación p/Amortizaciones de Inmovilizado
DT	Deuda Total = Pasivo Exigible = PT - FP = Fondos Ajenos = ELP + PC
E	Existencias
ELP	Exigible Largo Plazo = Acreed.a LP + Prov.Riesgos/Gtos.LP + Ing.a Dist.Vs.E.
FP	Fondos Propios = Patrimonio Neto
GDVS	Gastos a Distribuir en Varios Ejercicios
GF	Gastos Financieros
IF	Inmovilizado Financiero
IFT	Inversiones Financieras Temporales
IM	Inmovilizado Material
INTÉRVS/CRÉD.	Intérvalo sin Crédito = (AC - E - PC) / (Gtos.Explot. - DA - Provis.de Tráfico)
OING	Otros Ingresos
PC	Pasivo Circulante + Prov.Riesgos/Gtos.CP
PT	Pasivo Total = FP + ELP + PC
RA	Resultado de la Actividad = Rtdo. de las Actividades Ordinarias
RAIT	Rtdo. antes de Int. e Imp. = RAT + GF
RAIT y Ex.	Rtdo. antes de Int., Imp.y Extraordinarios = Rtdo. Actividad + GF
RAT	Rtdo. antes de Impuestos
RE	Rtdo. de la Explotación = Ingresos Exp. – Gastos Exp.
RF	Rtdo. Financiero
RN	Rtdo. Neto = Rtdo. después de Impuestos = Rtdo. del Ejercicio
RP	Recursos Permanentes = Pasivo Fijo = FP + ELP
T	Tesorería = Disponible
V	Ventas = Importe Neto de la Cifra de Negocios
VA	Valor Añadido = Ing. Exp. – Cons. Exp. - Otros Gtos. Exp. = RBE = Rtdo. Bruto de Exp.
	Consumos Exp. = Reducc. Exist. Prod. + APROV + GP
VM	Valor de Mercado. Cap. Social = Valor Bursátil de la Empresa = Valor de la Acción a FCE por cant. Acciones

El siguiente paso es proceder a analizar la normalidad de los ratios a través de pruebas paramétricas y no paramétricas. También se verificó la existencia de diferencias significativas en las medias de los ratios de las empresas fracasadas y las sanas por medio del análisis de la varianza, lo que arrojó un total de 40 ratios que rechazan la hipótesis nula de igualdad de medias.

También se procedió a analizar las correlaciones entre los ratios mediante el coeficiente de asociación de Pearson y se aplicó el análisis factorial a fin de identificar cuáles son los ratios que captan mayor información y así reducir el número de ellos en la elaboración de los modelos multivariantes, lo que contribuirá también a interpretar las relaciones entre las variables.

Como resultado de la aplicación de la técnica factorial, habiéndose rotado los factores en forma ortogonal a través del método Varimax, el programa retuvo 14 componentes principales de las 72 variables introducidas. Con ellos se retiene el 93,17% de la varianza total y a partir del siguiente componente decae bastante el porcentaje de varianza retenido.

La Tabla 8 muestra los ratios que componen cada factor, en función de la mayor correlación con dicha componente y con coeficientes de correlación superiores a 0,5, presentándose también el porcentaje de la varianza retenido por cada componente y el acumulado.

Tabla 8: Ratios que componen cada factor utilizados en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

Factor	Ratios	%Var.	% Acumulado
Factor 1	R16, R17, R23, R18, R21, R19, R20, R22, R26, R27, R28, R31, R30, R29, R72	27,76	27,76
Factor 2	R01, R02, R63, R54, R60, 53, R57, R58, R59, R03, R15, R66	14,88	42,64
Factor 3	R49, R39, R32, R33, R50, R46, R51, R48	8,74	51,38
Factor 4	R08, R40, R47, R34, R68, R07, R35	7,79	59,17
Factor 5	R11, R12, R61, R52, R42, R56, R62	6,75	65,92
Factor 6	R43, R55, R24, R25	5,22	71,13
Factor 7	R06, R14, R04	4,13	75,27
Factor 8	R41, R45, R44	3,56	78,83
Factor 9	R65, R64	3,29	82,12
Factor 10	R05	2,98	85,10
Factor 11	R38, R67	2,35	87,46
Factor 12	R70, R69	2,19	89,65
Factor 13	R73, R09	1,81	91,46
Factor 14	R10, R13	1,71	93,17

4.2.3. Técnica estadísticas multivariantes utilizadas

El estudio empírico realizado a través del análisis multivariante se encuadra dentro de los problemas de generalización, en los que a partir de unos datos el modelo extrae una función matemática, que al suministrar otros datos el modelo da un output. Serrano (1995, p. 96) señala que la clasificación y la predicción son dos problemas típicos de generalización.

En los problemas de clasificación el output pertenece a un conjunto finito de clases, mientras que en los problemas de predicción no lo es.

En este estudio se intentará abarcar ambos problemas, para lo cual se estimará un modelo matemático que permita clasificar a las empresas que fracasan frente a las que no lo hacen y conocer cuáles son las características diferenciales de ambos tipos de empresas. Luego se avanza buscando resolver el segundo problema, el de predicción del fracaso empresarial, para lo cual se procederá a validar el modelo a fin de verificar la eficacia de su generalización.

Entre las técnicas estadísticas válidas para explicar y predecir la insolvencia empresarial se ha elegido utilizar el análisis discriminante y la regresión logística.

4.2.3.1. Análisis discriminante

El análisis discriminante lineal aplicado a la predicción de quiebra por Altman (1968), Deakin (1972), Edmister (1972), Blum (1974), Taffler (1974), etc. busca establecer, a partir de los ratios financieros calculados en base a los estados financieros, una función lineal que clasifique con el mayor grado de acierto a los dos grupos en que se divide la población, empresas fracasadas y empresas sanas.

La discriminación se lleva a cabo estableciendo las ponderaciones del valor teórico para cada variable de tal forma que maximicen la varianza entre-grupos frente a la varianza intra-grupos. La combinación lineal para el análisis discriminante, también conocida como función discriminante, se deriva de una ecuación que adopta la siguiente forma:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + W_3 X_{3k} + \cdots . + W_n X_{nk}$$

dónde:

$Z_{jk} = puntuación\ Z\ discriminente\ de\ la\ función\ discriminante\ j\ para\ el\ objeto\ k$

$a = constante$

$W_i = ponderación\ discriminante\ para\ la\ variable\ independiente\ i$

$X_{ik} = variable\ independiente\ i\ para\ el\ objeto\ k$

De esa forma si la puntuación Z de cada empresa, es mayor que cero, la misma se considera como sana y para Z menor que cero como fracasada.

El análisis discriminante exige una serie de supuestos: (1) normalidad de las variables independientes; (2) matrices de covarianzas iguales para los dos grupos; (3) distribuciones independientes de ambos grupos y sin solapamiento entre ambas.

El primero de ellos ha sido investigado por autores como Deakin (1976) y Frecka y Hopwood (1983) quienes verifican una falta de normalidad en casi todos los ratios

analizados y un progresivo acercamiento a la distribución deseada cuando se les aplican transformaciones, se extraen los valores extremos o se realiza un acotamiento sectorial. Hair *et al* (1999, p. 264) comentan que existe evidencia contradictoria sobre la sensibilidad del análisis discriminante a incumplimientos del supuesto de normalidad multivariante de las variables independientes y de matrices de covarianza y dispersión desconocidas, pero iguales, para los grupos.

La estimación de la función discriminante se puede llevar a cabo utilizando el método de cálculo simultáneo o el método por etapas. El primero implica el cálculo de la función discriminante donde todas las variables independientes son consideradas simultáneamente, sin considerar la capacidad discriminante de cada variable. Este método se considera apropiado cuando, por razones teóricas, se quiere introducir todas las variables independientes en el análisis y no se está interesado en observar resultados intermedios basados solamente en las variables que discriminan mejor.

La estimación por etapas es una alternativa al enfoque simultáneo, el cual incluye las variables independientes dentro de la función discriminante de una en una, según su capacidad discriminatoria, siendo útil este método cuando se quiere considerar un número relativamente grande de variables independientes para incluir en la función, por lo que seleccionando secuencialmente la siguiente variable que mejor discrimina en cada paso, se eliminan las variables que no son útiles para discriminar entre los grupos y se identifica un conjunto reducido de variables. Este conjunto reducido es generalmente tan bueno como el conjunto completo de variables.

4.2.3.2. Análisis Logit

El análisis logit es una técnica de probabilidad condicional que se utiliza para estudiar la relación entre una serie de características de un individuo y la probabilidad de que dicho individuo pertenezca a uno de entre los dos grupos establecidos a priori. Esta técnica fue aplicada por autores como Ohlson (1980), Gentry *et al* (1985) y Casey y Bartczak (1984) y (1985) en la predicción de insolvencia de empresas industriales.

La variable dependiente es cualitativa, de carácter dicotómico, en donde el valor 0 indica en este estudio que la empresa es fracasada y el valor 1 que la empresa es sana.

La regresión logística tiene la ventaja de verse menos afectada que el análisis discriminante cuando no se cumplen los supuestos básicos, ya que no plantea restricciones respecto a la normalidad en la distribución de las variables independientes, ni respecto a la igualdad de matrices de varianzas y covarianzas de cada grupo. El modelo logit se basa en la función de distribución logística:

$$P_i = F(z_i) = \frac{1}{1 + e^{-zi}}$$

Siendo Z_i una combinación lineal de una o más variables independientes, en la cual los β representan los coeficientes a estimar:

$$Z_i = \beta_0 + \beta_1 x_i + \cdots \ldots + \beta_k X_K$$

Para este estudio se utiliza el paquete estadístico SPSS 10.0 que arroja los coeficientes de la función Z_i por lo que para aplicar el modelo se debe colocar dicha función lineal dentro de la función de distribución logística presentada. En la ecuación de regresión, la variable dependiente (Z_i) es el logaritmo del cociente entre la probabilidad de que una empresa sea fracasada y su complementario, la probabilidad de que sea sana. El análisis logit transforma el problema de predecir probabilidades comprendidas entre 0 y 1 en el problema de pronosticar una variable (P_i / 1 – P_i) que puede tomar cualquier valor real.

Dado que el modelo proporciona un valor continuo de probabilidad de respuesta entre 0 y 1, se ha utilizado como punto de corte una probabilidad estándar de P(F) = 0,5, a efectos de compararla con la probabilidad obtenida y así proceder a clasificar cada una de las observaciones como fracasada o sana.

4.2.3.3. Validación de los modelos elaborados

La validación de los resultados de los modelos de predicción de insolvencia con una muestra diferente de la de estimación es necesaria para otorgarle a los mismos poder predictivo. Normalmente se subdivide la muestra en dos partes de modo a obtener una muestra de estimación y otra de validación.

Por estar conformada la muestra de empresas fracasadas por 30 empresas se ha preferido no segmentarla con fines de validación adoptando otro recurso que es la validación de los modelos con la misma muestra de estimación pero dejando un elemento afuera, el cual fue llamado por Lachenbruch y Mickey (1968, p. 5) de “Método U”.

El método utilizado para validar los modelos, propuesto por Lachenbruch (1967) consiste en que la primera muestra se compone de todos los casos a excepción del primero, que sirve para realizar la validación. La segunda muestra está formada por todos los casos excepto el segundo, con el que se efectuará el test, y así sucesivamente. Luego se computan los errores Tipo I y Tipo II de los test realizados con cada uno de los casos. De esa forma, se consigue validar el modelo estimado sin precisar de una muestra diferente.

Dambolena y Khoury (1980, p. 1024), que utilizaron el método de validación “dejar uno afuera”, propuesto por Lachenbruch, para validar el modelo de predicción de fracaso estimado, sostienen que este procedimiento es ampliamente aceptado como el mejor método de validación a menos que la muestra sea muy grande, en cuyo caso el método clásico de validación es normalmente utilizado.

4.2.4. Resultados del análisis discriminante

La función discriminante del modelo elegido para el primer año anterior al fracaso la componen los ratios: R20, R36 y R56, los cuales fueron determinados en tres etapas, habiéndose incorporado en ese orden.

$$Z = 0{,}583 + 7{,}937 \cdot R20 + 0{,}352 \cdot R36 - 2{,}403 \cdot R56$$

Esta función resultante alcanza un porcentaje de acierto global de 93,3%.

El ratio R20 de rentabilidad, RAIT y Ext./AT, definido como resultado antes de intereses, impuestos y extraordinarios a activo total, es el que más contribuye en la discriminación de los grupos. Este ratio compone el Factor 1 del análisis factorial, con un coeficiente de correlación con la componente principal de 0,914. Dicho factor lo integran un conjunto de ratios de rentabilidad económica y otro de rentabilidad sobre componentes del pasivo. Su media muestra que las empresas fracasadas presentan rentabilidad económica negativa del 17% frente a un 4% positivo de las empresas sanas.

El ratio R36, también de rentabilidad, CFT/RE, determinado por la relación del cash flow tradicional a resultados de la explotación, fue seleccionado en la segunda etapa del proceso, no presentando correlación alta con ninguna componente principal. La media de este ratio de generación de recursos sobre los resultados de explotación en las empresas fracasadas es de 66% en cuanto en las sanas es de 145%.

El ratio R56 de endeudamiento, ELP/AT, definido como exigible a largo plazo a activo o pasivo total, fue el último a incorporarse a la función, y hace parte del Factor 5 que lo integran ratios de estructura del activo y del pasivo, presentando una correlación con el componente de 0,773. La media de este ratio de endeudamiento a largo plazo en relación al pasivo total es de 24% en las empresas fracasadas y de 13% en las sanas.

La función discriminante del modelo para el segundo año anterior al fracaso quedó constituida de la siguiente manera:

$$Z = 0{,}213 + 14{,}672 \cdot R20$$

Esta función resultante alcanza un porcentaje de acierto global del 80%.

Los resultados obtenidos en la clasificación, tanto los relativos a la muestra global como los de error tipo I y tipo II, para los modelos de análisis discriminante estimados para n-1 y para n-2 son presentados en la Tabla 9

Tabla 9: Resultados del Análisis Discriminante en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

		Modelo n-1	Modelo n-2
Global	% Acierto	93,3	80,0
	% Error	6,7	20,0
Tipo I	% Acierto	90,0	80,0
	% Error	10,0	20,0
Tipo II	% Acierto	96,7	80,0
	% Error	3,3	20,0

4.2.5. Resultados del Análisis Logit

En la estimación de los modelos multivariantes a través del análisis logit, para los dos años anteriores al fracaso, se utilizó el método por pasos hacia delante, a través de la razón de verosimilitud, de tal forma que en el modelo final sólo queden aquellos ratios realmente significativos, siendo, por tanto, el valor crítico para seleccionar una variable 0,05 y para eliminarla 0,1. Tanto en el modelo para el año anterior al fracaso como en el del segundo año previo fueron incorporados, como variables independientes, los 70 ratios elaborados.

Los ratios que finalmente entran en el modelo para el primer año anterior al fracaso son R06, R20, R23 y Constante.

La mejora en el porcentaje de clasificación correcta de las empresas de la muestra, conforme se van introduciendo iterativamente nuevos ratios en el modelo, se puede observar a través de la Tabla 10. En el último paso, se alcanza un porcentaje de éxito en la clasificación del 95%, algo superior al análisis discriminante de 93,3%.

Tabla 10: Resultados del Análisis Logit para el año anterior al fracaso en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

Paso	Variable	-2 log verosim.	Mejora -2 log verosim.	Clasific. global correcta
0	Constante	83,178	-	50,0%
1	R20	27,835	55,343	91,7%
2	R23	20,547	7,288	91,7%
3	R06	14,517	6,03	95,0%

Los resultados alcanzados con el modelo logit dos años antes del fracaso derivado a través del método por pasos denotan la incorporación de un único ratio como variable independiente, además de la constante.

Inclusive este ratio, fue el R20 de rentabilidad, definido como resultado antes de intereses, impuestos y extraordinarios a activo total, que fue también el único elegido por el análisis discriminante para componer la función del segundo año anterior a la crisis, por lo que se remite a los comentarios ya efectuados sobre los descriptivos de este ratio expuestos en la Tabla 10.

Los resultados del modelo propuesto para el segundo año anterior al fracaso, expuestos en la Tablas 11, reflejan un porcentaje de acierto global de 81,4%, levemente superior al del análisis discriminante de 80,0%.

Tabla 11: Resultados del Análisis Logit para el segundo año anterior al fracaso en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

Paso	Variable	-2 log verosim.	Mejora -2 log verosim.	Clasific. global correcta
0	Constante	81,774	-	50,8%
1	R20	46,864	34,910	81,4%

Los resultados obtenidos en la clasificación, tanto los relativos a la muestra global como los de error tipo I y tipo II, para los modelos de análisis logit estimados para n-1 y para n-2 son presentados en la Tabla 12.

Tabla 12: Resultados del Análisis Logit en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

		Modelo n-1	Modelo n-2
Global	% Acierto	95,0	81,4
	% Error	5,0	18,6
Tipo I	% Acierto	93,3	79,3
	% Error	6,7	20,7
Tipo II	% Acierto	96,7	83,3
	% Error	3,3	16,7

Analizando la desagregación del tipo de errores de clasificación obtenidos, cabe destacar respecto al análisis discriminante, que el error de tipo I, que es el más grave,

con el logit es más bajo en el modelo para n-1, 6,7% versus 10% del análisis discriminante, en cuanto en los modelos para n-2 la mejora del logit se ha producido en el porcentaje de clasificación de las empresas sanas, que pasó de 80,0% en el discriminante para 83,3% con el logit.

4.2.6. Validación de los modelos

La validación de los modelos elaborados en este estudio, a través del Método U propuesto por Lachenbruch (1967), arrojó resultados bastante satisfactorios en los dos años previos al fracaso. Como puede observarse en la Tabla 13, los modelos elaborados para el año anterior al fracaso, a través del análisis discriminante y análisis logit, presentaron resultados de validación buenos, siendo el porcentaje de acierto global del discriminante de 92% y del logit de 83%.

Tabla 13: Resultados de la validación del modelo para año (n-1) en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

	Anál. Discriminante n-1		Análisis Logit n-1	
Acierto	Estimación	Validación	Estimación	Validación
Global	93,3	91,7	95,0	83,3
Tipo I	90,0	86,7	93,3	83,3
Tipo II	96,7	96,7	96,7	83,3

La validación de los modelos estimados para el segundo año anterior al fracaso, expuestos sus resultados en la Tabla 14, también resultó muy buena en ambos modelos, lo que permite considerar los modelos estimados válidos para predecir la insolvencia de otras empresas de similares características a las de la muestra utilizada.

Tabla 14: Resultados de la validación del modelo para año (n-2) en el Estudio de Insolvencia de Empresas Cotizadas Españolas Pinheiro y Pinheiro

	Anál. Discriminante n-2		Análisis Logit n-2	
Acierto	Estimación	Validación	Estimación	Validación
Global	80,0	78,4	81,4	78,3
Tipo I	80,0	80,0	79,3	76,7
Tipo II	80,0	76,7	83,3	80,0

4.2.7. Conclusiones

Los modelos multivariantes estimados mediante ambas técnicas incorporaron tres ratios cada uno en el año anterior al fracaso, siendo éstos de rentabilidad, endeudamiento y liquidez, lo que demuestra la mejor capacidad discriminante de estos ratios, y en dicho orden.

El ratio más importante en los modelos para el año anterior al fracaso, inclusive ha sido el mismo con ambas técnicas estadísticas, corresponde a un ratio de rentabilidad de la inversión definido como resultado antes de intereses, impuestos y extraordinarios a activo total. Los segundos ratios que intervienen en las funciones son también de rentabilidad, siendo en el análisis discriminante el ratio de generación de recursos a resultado de la explotación y en el análisis logit el de generación de recursos a activo total, y el tercer ratio que interviene en cada una de las funciones es en el análisis discriminante el ratio de endeudamiento a largo plazo y en el análisis logit el de posición de efectivo.

Estas variables explicativas del fracaso, reflejan que el perfil de las empresas que acabaron suspendiendo pagos o quebrando estaban en los dos años previos con rentabilidad económica fuertemente negativa, con generación de recursos en relación al resultado de explotación inferior al 50% del de las empresas que continuaron normalmente su actividad y cuyo endeudamiento a largo plazo es prácticamente el doble que el de las empresas sanas.

Los modelos elaborados para el primer año anterior al fracaso presentan un porcentaje de acierto global en torno al 95% y en el segundo año anterior al fracaso éste se sitúa en el 80%. El modelo logit reveló resultados de clasificación levemente superiores a los del análisis discriminante en el año anterior al fracaso y similares en el segundo año, sin embargo, en la validación de los modelos el análisis discriminante se mostró mejor que el logit, manteniéndose en aquel los porcentajes de acierto de la estimación en los modelos de los dos años, en cuanto en el análisis logit el porcentaje de acierto global del año anterior al fracaso decayó bastante, pasando del 95% al 83,3%.

4.3. Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

En este estudio Zuleyka Díaz Martínez , José Fernández Menéndez y Mª Jesús Segovia Vargas (2004) realizan una investigación de carácter empírico consistente en la aplicación al sector de empresas de seguros no-vida del algoritmo de inducción de reglas y árboles de decisión See5, desarrollado por Quinlan (1997), a partir de un conjunto de ratios financieros de una muestra de empresas españolas de seguros no-vida, con el objeto de comprobar su utilidad para la predicción de insolvencias en este sector. También se comparan los resultados alcanzados con los que se obtienen aplicando la metodología Rough Set. Estas técnicas, procedentes del campo de la Inteligencia Artificial, no presentan los problemas que dificultan su aplicación en el ámbito empresarial, ya que, generalmente, se trata de modelos basados en una serie de hipótesis sobre las variables explicativas que en muchos casos no se cumplen y, además, dada su complejidad, puede resultar difícil extraer conclusiones de sus resultados para un usuario poco familiarizado con la técnica.

4.3.1. Las técnicas de clasificación

Dentro del conjunto de técnicas a las que se puede recurrir si se pretende analizar la información contenida en cualquier base de datos de tipo contable, destacan las de clasificación, en las que, a partir de un conjunto de variables explicativas, se trata de predecir la pertenencia de la empresa considerada a una serie de grupos o categorías mutuamente excluyentes. Los sistemas clasificadores proceden fundamentalmente de la Estadística o de la Inteligencia Artificial y, entre otras, una manera de diferenciarlos radica en la forma en la que se construyen. Así, se puede entender la construcción de un sistema clasificador con n variables explicativas como la división del espacio n-dimensional que forman esas variables en una serie de regiones, cada una de las cuales se asigna a una de las categorías previamente definidas. Esta división puede ser realizada de dos formas:

- Definiendo una o varias hipersuperficies separadoras. Algunas técnicas estadísticas muy empleadas en el análisis de la solvencia empresarial siguen este enfoque (discriminante, logit, probit).

- Realizando particiones sucesivas en el espacio de las variables explicativas, empleando una sola variable en cada partición. Dentro de este segundo tipo de sistemas clasificadores los Sistemas Expertos son la rama de la Inteligencia Artificial más empleada en la gestión empresarial, si bien sus posibilidades se han visto enriquecidas con los nuevos enfoques que han aportado otras técnicas de aparición más reciente, como los algoritmos de inducción de reglas y árboles de decisión, entre los que se encuentra el See5.

4.3.2. El algoritmo de inducción de reglas y árboles de decisión See5

Encuadrado dentro de las técnicas de Aprendizaje Automático (Machine Learning), el algoritmo See5 (Este algoritmo constituye una extensión de los algoritmos ID3 y C4.5. Una descripción detallada puede verse en Quinlan, 1993, 1997), permite construir automáticamente a partir de un conjunto de datos de entrenamiento un árbol de clasificación. Para inferir el árbol, el algoritmo realiza particiones binarias sucesivas en el espacio de las variables explicativas, de forma que en cada partición se escoge la variable que aporta más información en función de una medida de entropía o cantidad de información. El árbol así construido consta del mínimo número de atributos (variables) que se requieren para la clasificación correcta de los ejemplos dados, con lo que es claro el alto poder explicativo de esta técnica.

También se pueden elaborar, a partir del árbol, reglas de clasificación fácilmente interpretables, que definen las características que más diferencian a las distintas categorías establecidas inicialmente.

Este tipo de sistemas clasificadores presentan la ventaja, frente a las técnicas estadísticas, de que tienen un carácter estrictamente no paramétrico. Además, aunque no alcanzan el poder predictivo de las redes neuronales, sus resultados son mucho más fácilmente interpretables.

4.3.3. Análisis empírico: metodología y resultados.

La muestra de empresas seleccionada en este estudio, es la utilizada para la aplicación de la metodología Rough Set a la predicción de crisis empresariales en seguros no-vida en el trabajo de Segovia (2003). Consta de 36 empresas no fracasadas y 36 empresas fracasadas, emparejadas por tamaño y tipo de negocio, eliminando así el efecto de estas variables en el estudio, y escogiendo como criterio de selección de las empresas fracasadas el hecho de haber sido intervenidas por la Comisión Liquidadora de Entidades Aseguradoras (CLEA), por entender que se trata de una medida objetivamente determinable de las empresas que fracasan.

Una vez tomada la muestra, se sitúa en el periodo anterior al de la insolvencia para tratar de determinar qué indicios de este suceso proporcionan los datos de las cuentas anuales en forma de ratios. El éxito o fracaso de una empresa será entendido como una variable dependiente que deberá ser explicada por un conjunto de ratios financieros que actuarán como variables independientes. Así que de cada una de las empresas se han obtenido las cuentas anuales del año previo a la quiebra y, a partir de dicha información, se han calculado una serie de ratios, unos populares en la literatura contable para medir la solvencia empresarial y otros específicos del sector asegurador. En la Tabla 15 se presentan los 21 ratios seleccionados.

De las 72 empresas de que consta la muestra original, se han utilizado únicamente 27 empresas de cada una de las submuestras para la elaboración de los modelos, reservando las 9 restantes para poder comprobar la validez de los mismos aplicándolos a empresas cuyos datos no hubieran sido utilizados en dicha elaboración. En consecuencia, se tiene una muestra de entrenamiento para obtener los árboles y reglas de decisión formada por 54 empresas y una muestra de validación para verificar su capacidad predictiva formada por 18 empresas. El algoritmo se aplicó utilizando el programa See5.

Tabla 15: Ratios utilizados en el Estudio de Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

Ratio	Definición
R1	Fondo de Maniobra / Activo Total
R2	Beneficio antes de Impuestos(BAI)/ Capitales propios
R3	Ingresos Financieros/ Total Inversiones
R4	BAI*/ Pasivo Total BAI* = BAI+ Amortizaciones + Provisiones + Resultados Extraordinarios
R5	Total Primas adquiridas de seguro directo / Capitales propios
R6	Total Primas adquiridas de negocio neto / Capitales propios
R7	Total Primas adquiridas de seguro directo / Capitales propios + Provisiones Técnicas
R8	Total Primas adquiridas de negocio neto /Capitales propios + Provisiones Técnicas
R9	Capitales Propios / Pasivo Total
R10	Provisiones Técnicas / Capitales Propios
R11	Gastos Técnicos de seguro directo / Capitales propios
R12	Gastos Técnicos de negocio neto / Capitales propios
R13	Gastos Técnicos de seguro directo / Capitales propios + Prov. Técnicas
R14	Gastos Técnicos de negocio neto / Capitales propios + Provisiones Técnicas
R15	Ratio Combinado 1 = Ratio Siniestralidad de seguro directo (RSD)+ Ratio de Gastos (RG) RSD = Gastos Técnicos de seguro directo/ Total Primas adquiridas de seguro directo RG = Comisiones y otros gastos de explotación/ Otros ingresos explotación
R16	Ratio Combinado 2 = Ratio Siniestralidad de negocio neto (RSN)+ Ratio de Gastos (RG) RSN = Gastos Técnicos de negocio neto/ Total Primas adquiridas de negocio neto RG = Comisiones y otros gastos de explotación/ Otros ingresos explotación
R17	(Gastos Técnicos de seguro directo + Comisiones y otros gastos de Explotación)/ Total Primas adquiridas de seguro directo
R18	(Gastos Técnicos de negocio neto + Comisiones y otros gastos de Explotación)/ Total Primas adquiridas de negocio neto
R19	Provisiones Técnicas de reaseguro cedido / Provisiones Técnicas
R20	RSD = Gastos Técnicos de seguro directo/ Total Primas adquiridas de seguro directo
R21	RSN = Gastos Técnicos de negocio neto/ Total Primas adquiridas de negocio neto

El resultado obtenido de la aplicación del algoritmo See5 es el árbol de decisión que se muestra a continuación:

Ilustración 2: Árbol de decisión inicial obtenido de aplicación del algoritmo See5 en el Estudio de Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

```
R13 > 0.68:
:...R9 <= 0.59: mala (14)
:   R9 > 0.59:
:   :...R17 <= 0.99: mala (3)
:       R17 > 0.99: buena (3)
R13 <= 0.68:
:...R1 > 0.29: buena (20/2)
    R1 <= 0.29:
    :...R2 > 0.04: mala (3)
        R2 <= 0.04:
        :...R6 > 0.64: buena (3)
            R6 <= 0.64:
            :...R9 <= 0.85: mala (4)
                R9 > 0.85: buena (4/1)

Evaluation on training data (54 cases):

      Decision Tree
    ----------------
    Size      Errors

       8     3(5.6%)   <<


     (a)   (b)    <-classified as
    ----  ----
      27          (a): class buena
       3    24    (b): class mala


Evaluation on test data (18 cases):

      Decision Tree
    ----------------
    Size      Errors

       8    5(27.8%)   <<


     (a)   (b)    <-classified as
    ----  ----
       7     2    (a): class buena
       3     6    (b): class mala
```

En el árbol aparecen únicamente 6 de los 19 ratios iniciales, lo que indica que 13 de los ratios empleados no aportan información relevante para clasificar las empresas como

"buenas" o "malas". El árbol proporciona el menor número de ratios necesarios para alcanzar el objetivo deseado. El árbol se leería del modo siguiente:

- Si el ratio R13 es mayor de 0,68 y además el ratio R9 es menor o igual de 0,59, la empresa será "mala", siendo 14 el número de empresas de la muestra que verifican este hecho.
- Si el ratio R13 es mayor de 0,68 y el ratio R9 es mayor de 0,59 y el ratio R17 menor o igual de 0,99, la empresa será "mala", cumpliendo estas condiciones 3 empresas.

Y así se continúa descendiendo por el árbol, hasta completar un total de 8 hojas.

Al final de cada hoja aparece un valor (n) o (n/m): n representa el número de empresas en la muestra que se clasifican de acuerdo a las condiciones que nos llevan hasta esa hoja y m el número de empresas mal clasificadas.

La evaluación de este árbol de decisión construido con la muestra de entrenamiento (54 empresas) indica que el árbol consta de 8 ramas y comete un total de 3 errores (5,6%), lo que supone un porcentaje de aciertos del 94,4%. También se muestra una matriz de confusión que señala el tipo de errores cometidos.

Por último, para comprobar la capacidad predictiva del árbol, se clasifican de acuerdo con éste las 18 empresas de la muestra de validación, obteniendo un porcentaje de clasificaciones correctas del 72,2%.

See5 permite también establecer un conjunto de reglas más simples de la forma si (condiciones) - entonces (decisión). Las reglas que se obtienen a partir del árbol anterior son las siguientes:

Ilustración 3: Reglas de decisión obtenidas de aplicación del algoritmo See5 en el Estudio de Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

```
Rules:

Rule 1: (20/2, lift 1.7)
  R1 > 0.29
  R13 <= 0.68
  ->  class buena  [0.864]

Rule 2: (12/1, lift 1.7)
  R2 <= 0.04
  R6 > 0.64
  R13 <= 0.68
  ->  class buena  [0.857]

Rule 3: (7/1, lift 1.6)
  R9 > 0.85
  ->  class buena  [0.778]

Rule 4: (14, lift 1.9)
  R9 <= 0.59
  R13 > 0.68
  ->  class mala  [0.938]

Rule 5: (7, lift 1.8)
  R13 > 0.68
  R17 <= 0.99
  ->  class mala  [0.889]

Rule 6: (26/6, lift 1.5)
  R1 <= 0.29
  ->  class mala  [0.750]

Default class: buena
```

Cada regla consiste en:

- Una serie de estadísticas (n, lift x) o (n/m, lift x); n y m representan lo mismo que en el árbol y x es el resultado de dividir la precisión estimada de la regla entre la frecuencia relativa de la clase predicha en la muestra de entrenamiento. La precisión de la regla se estima mediante el denominado ratio de Laplace (n-m+1)/(n+2).
- Una o más condiciones que deben ser satisfechas para que la regla sea aplicable.

- La clase predicha por la regla.
- Un valor entre 0 y 1 que indica el nivel de confianza con el que ha sido hecha la predicción.

También existe una clase por defecto (en este caso “buena”) para cuando ninguna de las reglas sea aplicable.

El número de errores cometidos al clasificar mediante estas reglas y el tipo de los mismos coinciden, tanto con la muestra de entrenamiento como con la de validación, con los de las clasificaciones hechas con el árbol.

A pesar de que los resultados obtenidos son satisfactorios, es posible mejorarlos recurriendo a la opción que incorpora See5 de *adaptive boosting*, basado en el trabajo de Freund y Schapire (1997). La idea consiste en generar varios clasificadores (árboles o conjuntos de reglas) en vez de sólo uno. Como primer paso, se construye un único árbol (o conjunto de reglas) del mismo modo que se acaba de ver, que cometerá algunos errores en la clasificación (3 en este caso). Estos errores serán el foco de atención al construir el segundo clasificador en aras de corregirlos. En consecuencia, el segundo clasificador generalmente será diferente al primero y también cometerá errores que serán el foco de atención durante la construcción del tercer clasificador. Este proceso continúa para un número predeterminado de iteraciones o trials. Mediante este procedimiento, se consigue obtener un clasificador verdaderamente preciso. Así, partiendo del primer árbol de decisión los resultados que se alcanzan realizando 18 iteraciones son:

- Con la muestra de entrenamiento, el 100% de clasificaciones correctas, como se puede observar en la matriz de confusión:

Ilustración 4: Resultado de la aplicación del *adaptive boosting* del algoritmo See5 en el Estudio de Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

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

- Con la muestra de validación, el 83,3% de clasificaciones correctas:

Ilustración 5: Resultado de la validación del adaptive boosting del algoritmo See5 en el Estudio de Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

```
 (a)    (b)     <-classified as
----   ----
  7      2     (a): class buena
  1      8     (b): class mala
```

4.3.4. Enfoque mediante la Teoría Rough Set

La Teoría Rough Set fue propuesta por Pawlak a comienzos de los ochenta. Es un enfoque que se encuadra también dentro de las aplicaciones de la Inteligencia Artificial. Consiste en descubrir dependencias entre atributos en una tabla de información y reducir el conjunto de los mismos eliminando aquéllos que no son esenciales para caracterizar el conocimiento. Un reducto se define como el mínimo subconjunto de atributos que asegura la misma calidad de clasificación que el conjunto de todos ellos. De la tabla de información reducida pueden derivarse reglas de decisión en forma de sentencias lógicas (si <condiciones> entonces <decisión>). Los principales conceptos de esta teoría se exponen en Segovia (2003).

De la aplicación de la metodología Rough Set a la muestra de entrenamiento se obtuvieron 229 reductos (conjunto de ratios que clasifican a las 54 empresas en “buenas” y “malas” igual que si se tomaran en consideración los 19 ratios originales), de los cuales se seleccionó el formado por los ratios R3, R4, R9, R14 y R17 (Segovia, 2003).

Una vez seleccionado el reducto, se derivaron 27 reglas de decisión, que constituyen un algoritmo de clasificación que puede ser utilizado para evaluar cualquier otra empresa. El poder predictivo de este algoritmo se contrastó con la muestra de validación formada por 18 empresas.

Tal y como puede observarse en estas tablas, los resultados de ambas metodologías muestran su capacidad para responder de manera eficiente al problema de la

predicción del fracaso empresarial, siendo alternativas muy fiables a las técnicas estadísticas tradicionales, y más aún en el caso del algoritmo See5, que obtiene un porcentaje de aciertos superior con la muestra de validación clasificando correctamente el 88,89% de las empresas "malas" frente al 77,78% del Rough Set, lo cual es importante teniendo en cuenta que precisamente lo que interesa captar es la insolvencia.

Tabla 16: Resultados obtenidos de aplicación del algoritmo See5 frente al Rough Set en el Estudio de Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

See5

Clasificaciones correctas	Muestra de entrenamiento	Muestra de validación
Empresas "buenas"	100%	77,78%
Empresas "malas"	100%	88,89%
Total	100%	83,33%

Rough Set

Clasificaciones correctas	Muestra de entrenamiento	Muestra de validación
Empresas "buenas"	100%	77,78%
Empresas "malas"	100%	77,78%
Total	100%	77,78%

4.3.5. Conclusiones

Los dos métodos son estrictamente no paramétricos, lo que les convierte en superiores a las técnicas estadísticas en el sentido de que se adecúan más a la información contable, que suele presentar datos interrelacionados, incompletos, adulterados o erróneos; ofrecen productos muy sencillos entendibles fácilmente por el analista humano, ya sea en forma de árboles o reglas de decisión, realizando una clasificación de las empresas entre solventes e insolventes que permite determinar la importancia de cada variable en el proceso de asignación. Además, dan buenos resultados incluso cuando se trabaja con escaso número de datos, aspecto éste importante en las aplicaciones al ámbito financiero.

El algoritmo See5 trabaja mejor con datos discretos, bien sean variables cualitativas o variables cuantitativas previamente discretizadas, además acepta atributos de tipo discreto o continuo, sin ningún tipo de limitación.

La metodología Rough Set proporciona muchos subconjuntos, entre los cuales habrá de seleccionar el analista para derivar las reglas de decisión.

See5 posee una ventaja muy importante desde el punto de vista económico, ya que permite considerar distintos costes de clasificación errónea, mientras que Rough Set sólo computa el número global de errores sin distinguir si se trata de clasificar una empresa sana como fracasada o clasificar una fracasada como sana, error este último que resultaría mucho más grave.

4.4. La Inteligencia Artificial como una Alternativa Viable en la Predicción de la Insolvencia de Empresas de Seguros.

Un importante conjunto de procedimientos de Aprendizaje Automático es el formado por los distintos sistemas de inducción de árboles de decisión y conjuntos de reglas. Dentro de los estudios realizados siguiendo este enfoque, destaca el trabajo, con una orientación marcadamente estadística, de Friedman (1977), que sirvió como base para la construcción del sistema CART (*Classification and Regression Trees*), descrito en Breiman *et al.* (1984).

Más inspirados directamente en el campo de la Inteligencia Artificial son los algoritmos de inducción de árboles de decisión ID3, C4.5 y See5 desarrollados por Quinlan (1979, 1983, 1986, 1993, 1997), que han alcanzado una notable repercusión. En cuanto a su aplicación a la predicción de crisis en el sector de seguros español, Díaz Martínez *et al*. (2004) obtienen en su trabajo resultados muy superiores a los del Análisis Discriminante.

En la actualidad, como consecuencia fundamentalmente del proyecto comunitario denominado Solvencia II, la normativa reguladora del tratamiento de la solvencia de las entidades aseguradoras está sujeta a un proceso de revisión encaminado a conseguir un mayor ajuste a las circunstancias específicas de cada entidad de los requisitos en materia de solvencia a las que aquéllas 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.

Se han propuesto recientemente 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 se plantea en 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 se desarrolla 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 las empresas aseguradoras, siendo el fin último de esta investigación el desarrollar un modelo sencillo basado en ratios financieros de previsión de insolvencias en empresas de seguros no-vida, de tal forma que, en función del valor que alcancen en estas entidades el pequeño conjunto de ratios derivados en este modelo, se logre anticipar las insolvencias en este sector, orientando así los recursos limitados de la inspección hacia aquellas empresas preseleccionadas como posibles candidatas al fracaso.

Asimismo, se compararán los resultados alcanzados con los que se obtienen mediante la aplicación de Regresión Logística.

4.4.1. El Algoritmo PART

El problema que se plantea es un problema de clasificación: dado un conjunto de ratios financieros tratamos de clasificar a la empresa de la cual han sido obtenidos como

sana o fracasada. Aunque esta tarea puede ser llevada a cabo, y probablemente con notable éxito, por un experto humano, en lo que interesa es utilizar técnicas analíticas o algoritmos que permitan eliminar la subjetividad, y el coste, que supone la intervención de dicho experto humano. Es decir, interesa automatizar de algún modo el proceso de inferencia y toma de decisiones.

Para llevar a cabo esta labor ha surgido recientemente un amplio abanico de técnicas, muchas de ellas procedentes de las áreas de la Inteligencia Artificial y el Aprendizaje Automático y otras más estrechamente emparentadas con los métodos estadísticos clásicos. 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 poderosas herramientas que permiten generar 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 y exhiben una notable inteligibilidad que convierte su utilización en especialmente atractiva.

En este trabajo se utiliza 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, y se comparará con una herramienta estadística de clasificación de carácter estándar como es la Regresión Logística. 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. La estructura de condición y ramificación de un árbol de decisión es idónea para el problema que se tiene, el de clasificación. Debido al hecho de que la

clasificación trata con clases o etiquetas disjuntas, es decir, una instancia es de una clase o de otra, pero no de varias clases a la vez, un árbol de decisión conducirá un ejemplo hasta una y sólo una hoja, asignándole, por tanto, una única clase.

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 este 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 conduce 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. En cuanto al mecanismo de generación del 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 CART (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 se conoce 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 elementos, especialmente si se tiene 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 se está 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 pe, que será la tasa de error prevista o estimada. Entonces, para una hoja que cubra N casos, el número de errores previstos será N x pe . 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 of 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 le 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.

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.

El tratamiento de los valores desconocidos es similar al que lleva a cabo el algoritmo C4.5. 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.

4.4.2. Selección de datos y variables

La muestra de empresas que se ha utilizado en este 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. Dicha 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)1- y la otra por 36 empresas no fracasadas - que para los mismos periodos se mantenían en funcionamiento -, emparejadas por tamaño - medido a través del volumen de primas -, tipo de negocio y año de procedencia de los datos, eliminando así el efecto de estas variables en el estudio.

Una vez tomada la muestra, se va a periodos anteriores al de la insolvencia para tratar de determinar qué indicios de este suceso proporcionan los datos de las cuentas anuales en forma de ratios. El éxito o fracaso de una empresa será entendido como una variable dependiente que deberá ser explicada por un conjunto de ratios financieros que actuarán como variables independientes. Así que de cada una de las empresas se han obtenido las cuentas anuales de los dos años previos a la quiebra y, a partir de dicha información, se han calculado una serie de ratios, unos populares en la literatura contable para medir la solvencia empresarial y otros específicos del sector asegurador. En la tabla 17 se presentan los 25 ratios seleccionados.

Tabla 17: Ratios empleados en el estudio La Inteligencia Artificial como una alternativa viable.

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

Se van a desarrollar dos modelos diferentes según que los datos procedan de uno o dos años previos a la quiebra. De este modo, se trata de predecir la crisis con uno o dos años de antelación, respectivamente. Se llama a estos modelos Modelo 1 y Modelo 2.

Para desarrollar el Modelo 1, se utilizarán las 72 empresas disponibles. Sin embargo, al no disponer de los datos de la totalidad de estas empresas para el segundo año previo a la quiebra, únicamente de 70 de ellas. Al eliminar también las respectivas parejas de las dos empresas faltantes, se cuenta en total con 68 empresas para el desarrollo del Modelo 2.

Para verificar la capacidad predictiva de los modelos, se lleva a cabo un proceso de validación jackknife (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* (también denominado *leave-one-out*). Siendo k el número de instancias que contenga el conjunto de entrenamiento (en este caso, 72 para el Modelo 1 y 68 para el Modelo 2), 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 este, ofrece la oportunidad de conseguir la estimación más exacta que posiblemente pueda obtenerse.

Para la aplicación del algoritmo PART y la Regresión Logística a esta muestra se ha utilizado, respectivamente, el paquete gratuito de minería de datos WEKA desarrollado en la Universidad de Waikato (Witten y Frank, 2000) y el software R 2.1.0 distribuido gratuitamente por CRAN Foundation (R Development Core Team, 2005).

4.4.3. Resultados

Antes de pasar a comentar los resultados obtenidos con el algoritmo PART, se ha 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 0,3% 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%. Ya que se dispone de una muestra relativamente pequeña, se inclina por aprovechar los ejemplos que podrían ser desechados si se dispusiese de una gran cantidad de casos. Hay que pensar que en 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, se ha 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.

A continuación, se presenta la lista de decisión obtenida con PART para el primer año anterior a la quiebra, esto es, el Modelo 1. Al objeto de obtener un conjunto de reglas lo más general posible, de cara a la clasificación futura de nuevas empresas, se ha exigido que cada una de las reglas cubra al menos 11 empresas.

Ilustración 6: Modelo 1 en el estudio La Inteligencia Artificial como una alternativa viable.

Modelo 1

PART decision list

R3 > 1.186837 AND
R14 <= 0.67952: sana (40.8/10.8)

R3 <= 1.974321: fracasada (19.12/1.0)

: fracasada (12.08/5.0)

Number of Rules: 3

Como se puede observar, la lista de decisión consta de tan sólo tres 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 "sana". 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 "fracasada". Por último, si tampoco se verificase la segunda regla, a la empresa se le asignaría la clase por defecto ("fracasada"). Al final de cada una de las reglas de la lista se observan entre paréntesis unos valores 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. En este caso aparecen valores fraccionarios por el tratamiento de los valores perdidos que se ha comentado anteriormente.

Además de la indiscutible sencillez de la lista de decisión, ésta obtiene un porcentaje de acierto estimado mediante el método *jackknife* del 82%, resultado más que aceptable que genera confianza en la bondad del modelo. Este resultado se extrae de la siguiente matriz de confusión, donde se observa que se clasifican incorrectamente 8 de las 36 empresas fracasadas, lo que supone un error en la clasificación de las mismas del 22% - o acierto del 78% -, y 5 de las 36 empresas sanas, que equivale a un error del 14% o acierto del 86%, obteniéndose en global el porcentaje de acierto señalado del 82%.

Ilustración 7: Matriz de confusión modelo 1 en el estudio La Inteligencia Artificial como una alternativa viable.

```
=== Confusion Matrix ===
a   b  <-- classified as
28  8 | a = fracasada
5  31 | b = sana
```

Este Modelo 1 confirma la importancia de la liquidez de cara a predecir el fracaso empresarial, medida a través del ratio R3. Aunque la liquidez es una necesidad generalizada en cualquier tipo de empresa, en la empresa aseguradora dicha necesidad reviste una mayor importancia, ya que el proceso productivo es de sentido inverso al convencional (ésta cobra el importe de la prima antes de hacer frente al pago del siniestro u otra contraprestación) y, por tanto, no deberían presentarse problemas por falta de liquidez, sino por no haber invertido adecuadamente los recursos procedentes de las primas, puesto que por la propia naturaleza del negocio la empresa de seguros en funcionamiento normal dispondrá de liquidez permanente. La existencia de problemas de naturaleza financiera no significaría entonces una situación de suspensión de pagos sino directamente el aviso de una quiebra técnica, como consecuencia del mencionado ciclo inverso. Como señala Millán Aguilar (2000), en una empresa de seguros los problemas de liquidez deben aparecer con posterioridad a los problemas económicos y no al contrario, como ocurre en otros sectores.

Por otro lado, también se manifiesta la solvencia, medida a través del ratio R14, como factor determinante a la hora de predecir el fracaso en las empresas de seguros. Este

ratio 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, según la lista de decisión obtenida, cuando el activo circulante suponga más del 119% del pasivo circulante y los gastos técnicos no superen el 68% de la suma de fondos propios y provisiones técnicas, la empresa será considerada sana. Si esta regla no se cumpliese y, por tanto, la empresa no pudiese ser catalogada como sana, simplemente por tener un activo circulante igual o inferior al 197% del pasivo circulante la empresa sería considerada fracasada.

Con los datos del segundo año previo a la quiebra, se obtiene la siguiente lista de decisión, exigiendo que cada una de las reglas de la misma cubra al menos 8 empresas:

Ilustración 8: Modelo 2 en el estudio La Inteligencia Artificial como una alternativa viable.

Modelo 2

PART decision list

R3 > 0.912177 AND
R14 <= 0.616583: sana (40.0/11.0)

: fracasada (28.0/5.0)

Number of Rules: 2

El porcentaje de acierto estimado mediante el método *jackknife* que se extrae de la siguiente matriz de confusión es ahora del 72% (recordemos que para el año 2 cada una de las dos submuestras consta de 34 empresas). Como cabía esperar, se observa una disminución en la precisión clasificatoria ante el aumento del horizonte temporal de la predicción. No obstante, sigue siendo un resultado más que aceptable.

De nuevo se confirma la importancia de la liquidez y la solvencia a la hora de predecir el fracaso de las empresas aseguradoras, en esta ocasión con dos años de antelación. Así, cuando el activo circulante supere el 91% del pasivo circulante y los gastos técnicos no supongan más del 62% de la suma de provisiones técnicas y fondos propios, la empresa será clasificada como sana. En otro caso, se considerará fracasada.

Ilustración 9:Matriz de confusión modelo 2 en el estudio La Inteligencia Artificial como una alternativa viable.

=== Confusion Matrix ===

a	b	<-- classified as
23	11	a = fracasada
8	26	b = sana

4.4.4. Comparación con la 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 $log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots .. + \beta_p x_p + \varepsilon$ donde ε es el término de error y p la probabilidad de éxito en una variable aleatoria binaria que sigue una distribución de Bernoulli. Los valores que toma esta variable indican la clase a la que pertenece cada observación. Dada una nueva observación caracterizada por unos valores concretos de $x_1, x_2, \ldots\ldots, x_p$ el modelo 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, lo que obliga a la imputación de estos últimos si no se desea perder la información suministrada por los casos para los cuales algún atributo toma un valor desconocido (lo que no sería en absoluto recomendable con pequeños tamaños muestrales). Para llevar a cabo esta imputación se ha utilizado la técnica propuesta en Troyanskaya *et al*. (2001). En este artículo se comparan distintas estrategias de imputación para los missing values concluyéndose que la que proporciona mejores resultados es la denominada KNNimpute, que consiste en buscar para cada observación con algún missing value las observaciones más cercanas a ella (los "vecinos más próximos") con todos sus datos completos y estimar el missing value como una media ponderada de acuerdo con la distancia de los valores correspondientes de dichos vecinos más próximos.

Otro problema es el derivado de la necesidad de determinar cuáles serán las variables explicativas que se incluirán en el modelo eliminando aquéllas que resulten ser irrelevantes. 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 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 $-2log\left[L\left(\hat{\theta}\right)\right] + p \cdot logn$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). Los ratios seleccionados para cada modelo (12 para el primer año y 15 para el segundo), junto con los coeficientes estimados y su significatividad aparecen recogidos en el tabla 18.

Tabla 18: Coeficientes de los modelos de regresión logística en el estudio La Inteligencia Artificial como una alternativa viable.

Model 1:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.23398	1.20745	-0.194	0.84635	
R2	0.01571	0.01113	1.412	0.15797	
R4	0.45187	0.20085	2.250	0.02446	*
R6	-3.72381	1.75002	-2.128	0.03335	*
R7	4.79081	1.85247	2.586	0.00970	**
R8	1.99941	0.72062	2.775	0.00553	**
R10	-2.24256	1.04729	-2.141	0.03225	*
R11	3.09749	1.35527	2.286	0.02228	*
R13	-3.66636	1.34857	-2.719	0.00655	**
R15	2.05890	0.96345	2.137	0.03260	*
R19	-3.70863	1.65530	-2.240	0.02506	*
R24	8.05309	4.30867	1.869	0.06162	.
R25	-11.18083	5.18987	-2.154	0.03121	*

Model 2:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-18.86500	7.05863	-2.673	0.00753	**
R3	0.02748	0.01070	2.569	0.01019	*
R5	-3.31844	1.57387	-2.108	0.03499	*
R6	-57.45636	21.07848	-2.726	0.00641	**
R7	51.41808	19.22648	2.674	0.00749	**
R8	33.17818	13.00016	2.552	0.01071	*
R9	-27.73635	12.27665	-2.259	0.02387	*
R10	1.25591	0.66732	1.882	0.05983	.
R12	-44.38111	22.05338	-2.012	0.04417	*
R13	40.55367	22.46503	1.805	0.07104	.
R14	157.70892	91.71043	1.720	0.08550	.
R15	-157.27012	92.81663	-1.694	0.09019	.
R17	-71.56710	24.51203	-2.920	0.00350	**
R19	-73.55963	24.97920	-2.945	0.00323	**
R20	93.15534	30.39458	3.065	0.00218	**
R24	21.08603	7.64695	2.757	0.00583	**

---

*Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1*

El porcentaje de acierto en la clasificación estimado mediante el método *jackknife* es del 68% con el Modelo 1 y del 70,5% con el Modelo 2, porcentajes que se extraen de las siguientes matrices de confusión:

Ilustración 10: Matriz de confusión modelo 1 y 2 caso Regresión Logística en el estudio La Inteligencia Artificial como una alternativa viable.

Año 1

```
=== Confusion Matrix ===
  a   b   <-- classified as
 25  12 |  a = fracasada
 11  24 |  b = sana
```

Año 2

```
=== Confusion Matrix ===
  a   b   <-- classified as
 23   9 |  a = fracasada
 11  25 |  b = sana
```

4.4.5. Conclusiones

La detección precoz de la insolvencia empresarial es un problema que ha recibido constante atención por parte del mundo académico y profesional. El sector del seguro no ha sido ajeno a esta situación. La utilización de modelos eficientes de predicción de insolvencias facilitaría la labor de supervisión de las empresas aseguradoras permitiendo que los recursos limitados de la inspección se dirigiesen hacia aquéllas preseleccionadas como potencialmente insolventes y, de forma paralela, se flexibilizase la normativa en cuanto a requisitos de solvencia.

En este trabajo se ha aplicado a una muestra de empresas españolas de seguros no-vida, partiendo de un conjunto de ratios de carácter financiero, un paradigma procedente del área de la Inteligencia Artificial conocida como Aprendizaje Automático, el algoritmo de inducción de listas de decisión PART, con el objeto de comprobar su utilidad para la predicción de insolvencias en el mencionado sector. Al objeto de tener una referencia que pueda ser utilizada como término de comparación, se ha aplicado también Regresión Logística por ser ésta una técnica estadística estándar.

Los resultados obtenidos con PART mejoran a los que se alcanzan mediante Regresión Logística, especialmente en el primer año previo al fracaso. Además, el algoritmo PART supera a la Regresión Logística también en otros aspectos: se aplica con facilidad, proporciona modelos de interpretación más sencilla y es robusto ante el "ruido" introducido por valores faltantes y outliers, y por tanto se adecúa mejor a la información contable, que suele presentar datos interrelacionados, incompletos, adulterados o erróneos.

5. Desarrollo

Una de las áreas que ha tenido más estudio y avance en las últimas décadas, ha sido la Minería de Datos, debido principalmente al incremento en el tamaño de las Bases de Datos (BD), teniendo como resultado una falta de conocimiento de la información que se encuentra presente en ellas.

El objetivo principal de investigaciones sobre Minería de Datos, tratan de descubrir el conocimiento inmerso dentro de grandes BD.

Un ejemplo de estas BD, son las provistas por The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in DB 2007[10] (ECML PKDD 2007). ECML PKDD es uno de los congresos más importantes en el área de Aprendizaje Computacional (Machine Learning) y Minería de Datos, realizado en Europa cada año.

Minería de Datos es la exploración y análisis para la identificación no trivial de patrones (conocimiento) dentro de grandes cantidades de datos, que puedan ser válidos, novedosos, potencialmente útiles y entendibles. Los resultados que obtengamos al aplicar Minería de Datos deben de ser interesantes, siempre y cuando sean comprensibles (por seres humanos), sean válidos con cierto grado de certeza, sean potencialmente útiles y deben de ser novedosos para validar una hipótesis planteada, una vez que se tiene alguno de estos resultados se pueden evaluar de manera objetivamente (criterios estadísticos) o subjetivamente (perspectiva del usuario). Cuando se lleva a cabo una tarea de Minería de Datos, se puede describir en términos de:

a) Datos relevantes: Dentro de todos los datos que se tienen, se deben seleccionar aquellos que sean más importantes para analizar.
b) Conocimiento previo (Background knowledge): Hay que tener un conocimiento previo de la materia para que se pueda guiar el proceso de forma correcta.
c) Tipos de conocimiento: Se plantea el problema de encontrar un conocimiento sobre un conjunto de datos, por lo que es necesario establecer qué es lo que se

desea obtener, usualmente a esto se le conoce como la problemática asociado a la BD.

d) Medidas de rendimiento: Se deben de plantear estas medidas estadísticas para evaluar los resultados obtenidos.
e) Técnicas de representación: Siempre se debe tener establecida cuál va a ser la forma de representar los datos que estemos manejando.

La Minería de Datos es una fase importante para poder llevar a cabo el proceso de Extracción de conocimiento en Bases de Datos (Knowledge Discovery in Databases KDD). Áreas con las que conjunta la Minería de Datos:

1. Bases de Datos: Permite almacenar y gestionar grandes cantidades de datos que posteriormente permiten acceder a ellos de forma rápida y estructurada.
2. Estadística: Disciplina que va de la mano de la Minería de Datos y cuyos resultados se evalúan utilizando medidas estadísticas.
3. Inteligencia Artificial: de esta nacen los métodos de aprendizaje que nos van a permitir llevar a cabo una de las metas principales en Minería de Datos llamada clasificación.
4. Visualización: Con esta disciplina podemos presentar el conocimiento y los datos usados, de una manera más visual y comprensible.

Una de las tareas principales de KDD y Minería de Datos, es la clasificación. La clasificación es usada para predecir la clase de cada ejemplo presente en los datos y es realizada mediante diversos tipos de clasificadores. De acuerdo a J. Ross Quinlan, los siguientes tipos de clasificadores son los más utilizados para efectuar la clasificación de los datos.

a) Tablas de Decisión.
b) Reglas de Decisión.
c) Clasificadores Basados en Casos.
d) Redes Neuronales.
e) Clasificadores Bayesianos.

f) Clasificadores basados en Acoplamientos.

Nuestro trabajo estará enfocado a revisar detalladamente los clasificadores basados en árboles de decisión.

5.1. Base de datos SABI

La base de datos utilizada en el presente estudio va a ser la BD SABI - Sistema de Análisis de Balances Ibéricos. Posee un software avanzado que permite conocer y analizar los balances de más de 1,4 millones de empresas españolas y más de 400.000 portuguesas.

Permite análisis detallados, estadísticos y comparativos de empresas y grupos de empresas, así como la obtención de gráficos ilustrativos de los balances y cuentas de resultados. Ello facilita el seguimiento de la evolución financiera de las empresas en relación a sus competidores, así como los análisis del entorno de mercado/competencia (marketing) y la investigación económica en general.

Damos a continuación unos detalles a partir de unas imágenes para iniciarse en la utilización de esta base de datos.

Ilustración 11: SABI. Pantalla de entrada donde aparece el menú con las distintas posibilidades de búsqueda

Ilustración 12: SABI. Opción tipo de búsqueda por actividad con los diferentes códigos y subcódigos.

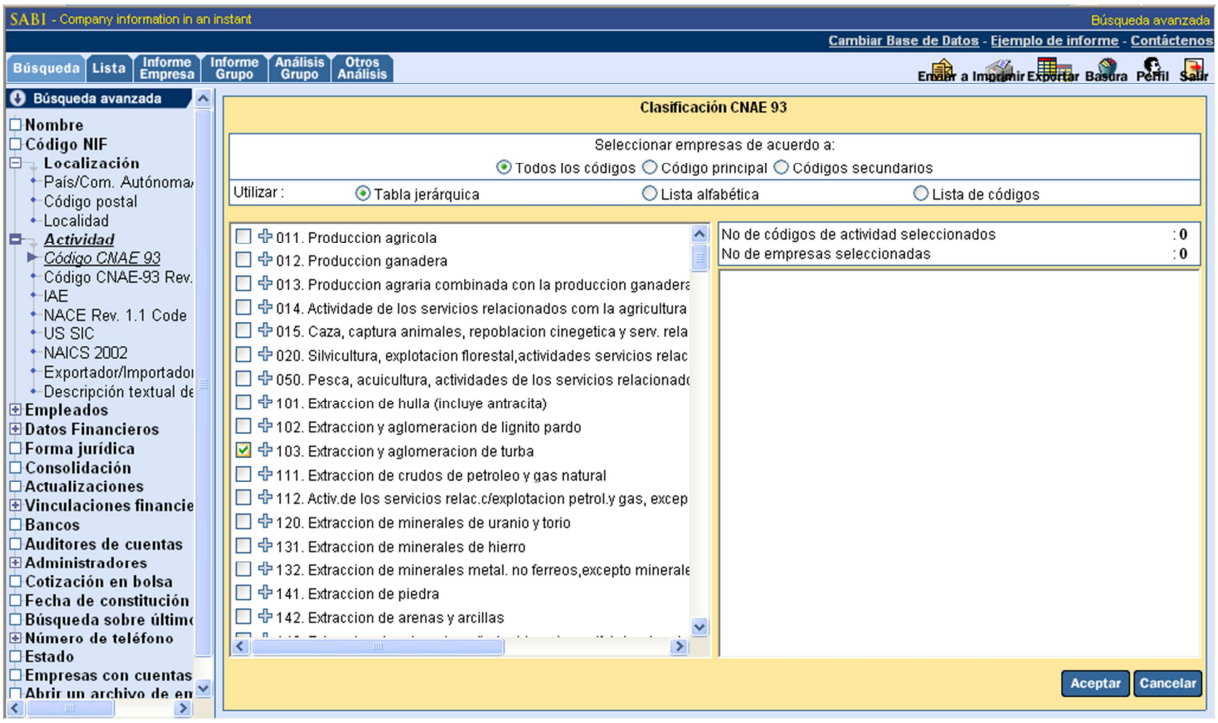


Ilustración 13: SABI. La opción del menú lista permite ver las empresas seleccionadas

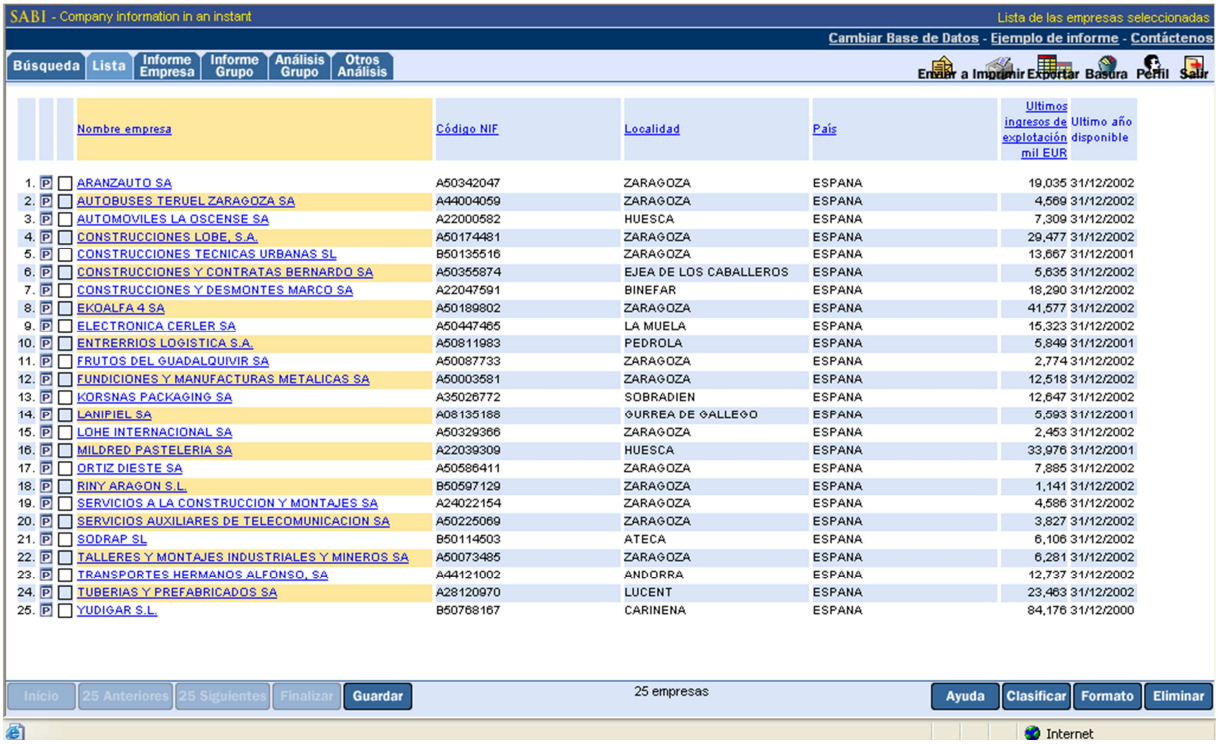


Ilustración 14: SABI. Existen distintas posibilidades de clasificación de la lista.

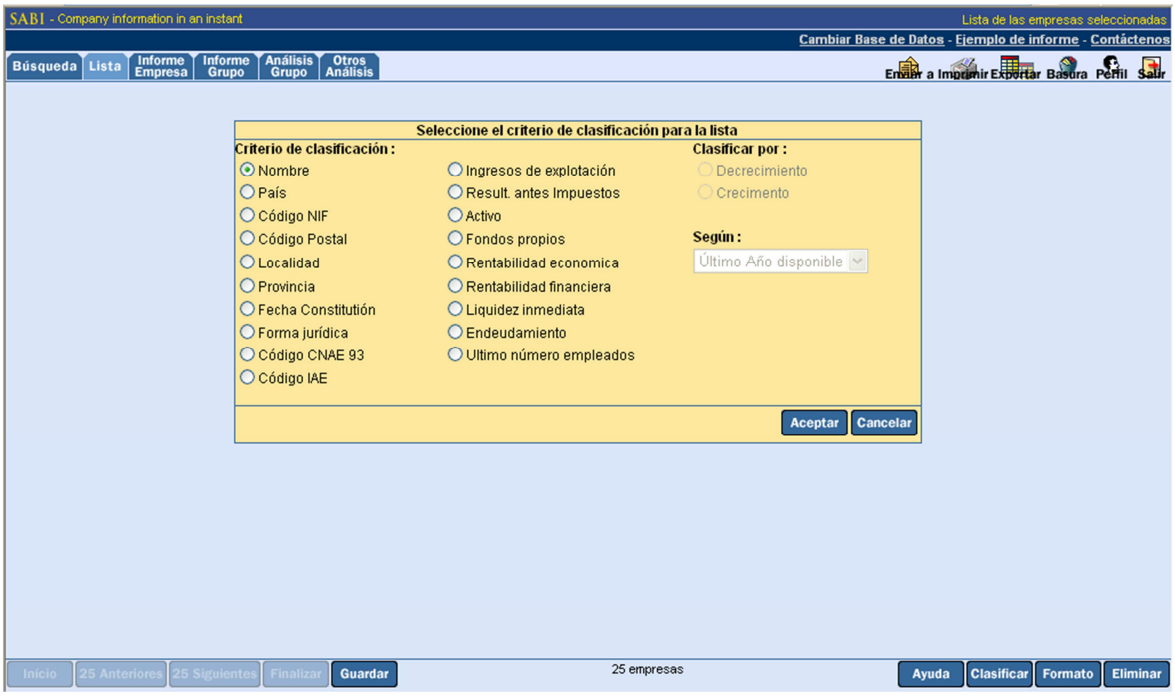


Ilustración 15: SABI. En el resto de pestañas aparecen Informe de empresa, Informe de grupo y otros análisis. El informe de empresa se compone de : Datos generales, perfil con las principales cuentas, balance, cuenta de pérdidas y ganancias, ratios, actividad, accionistas,

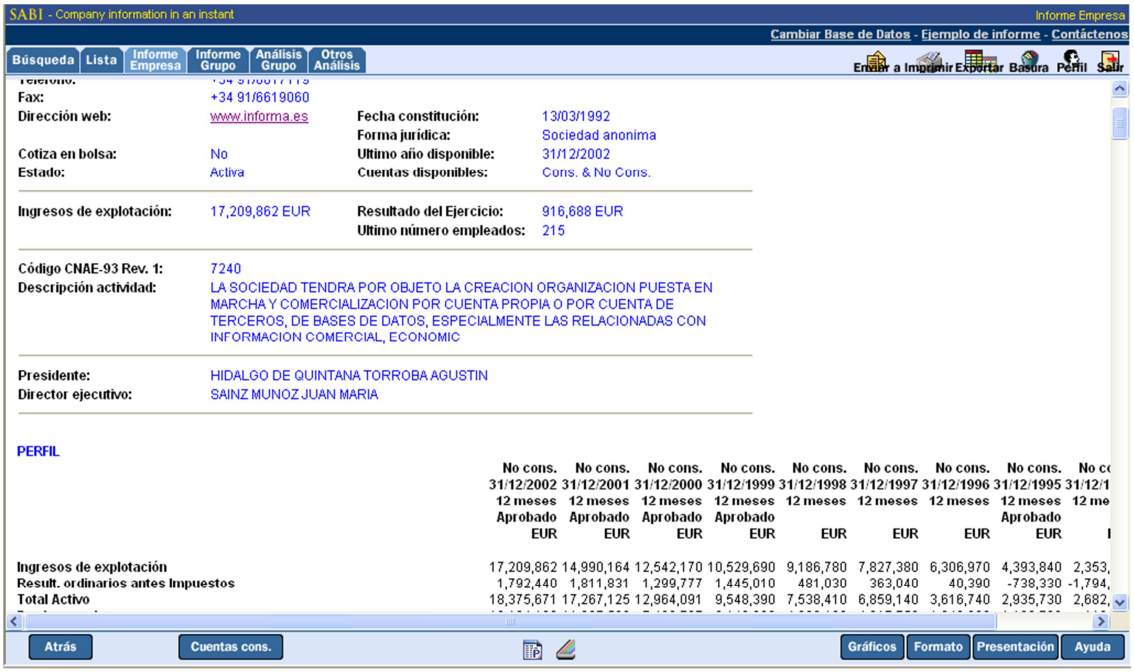


Ilustración 16: SABI. Ejemplo de gráfico

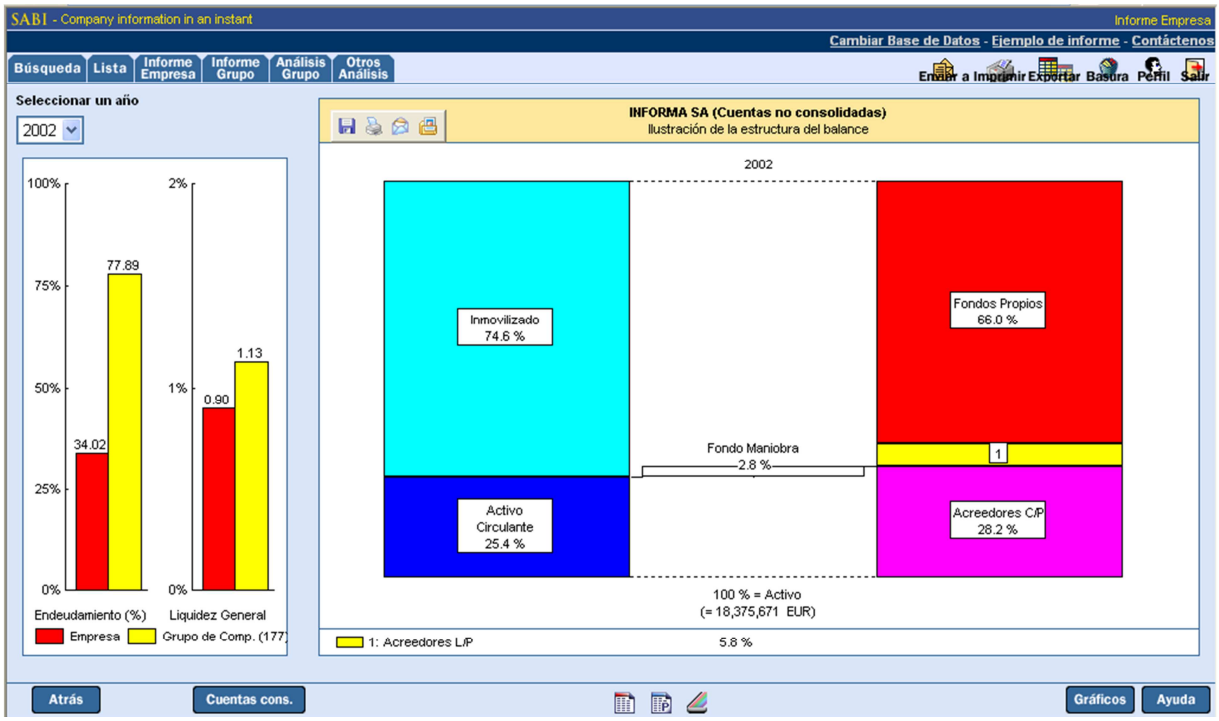


Ilustración 17: SABI. Evolución de índices

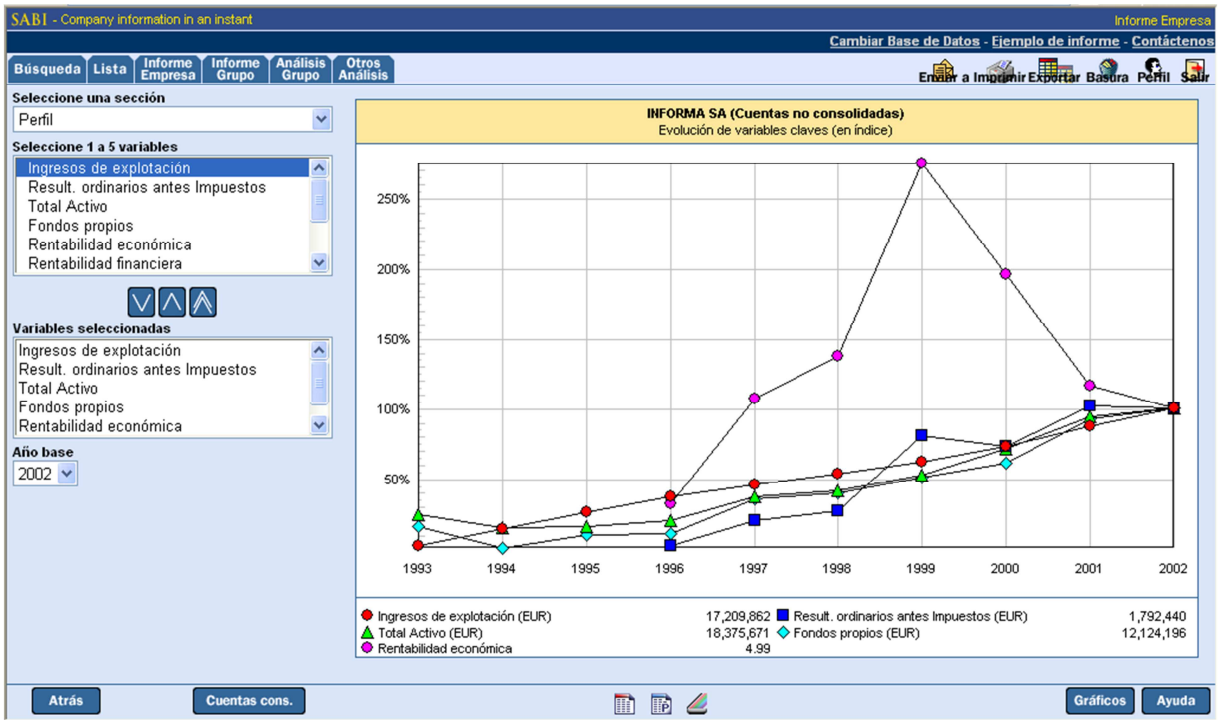


Ilustración 18: SABI. Comparativas con empresas del mismo sector y competidores

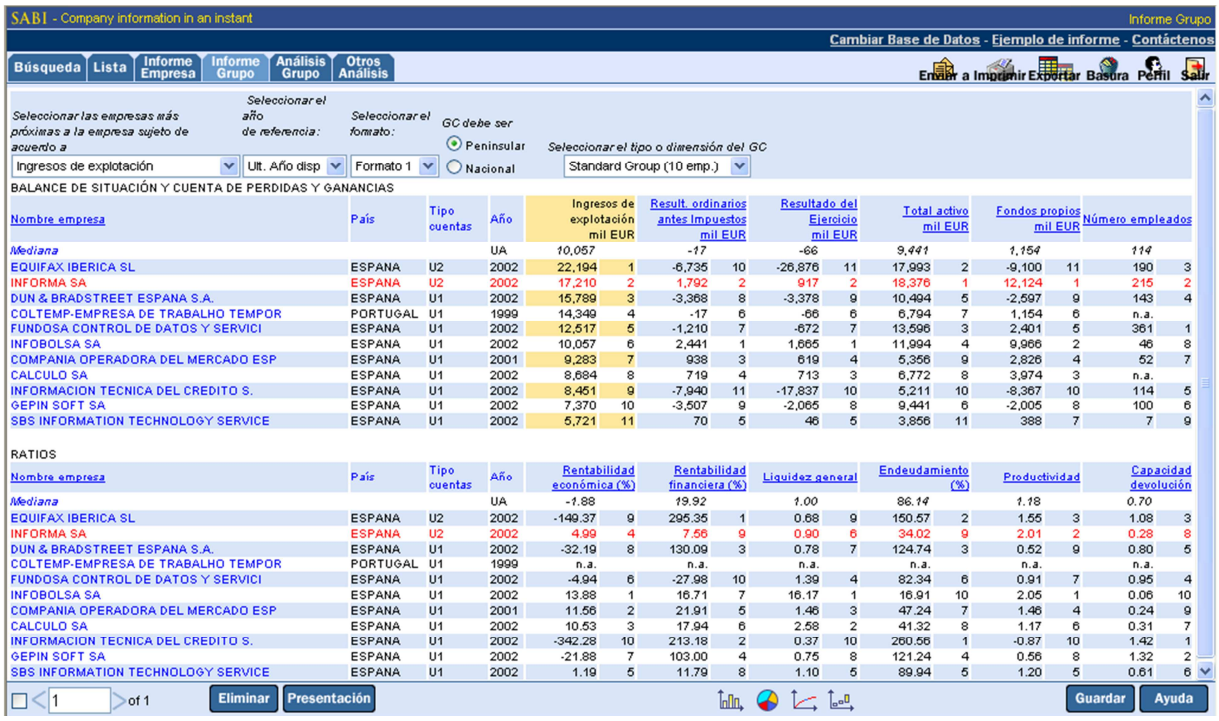
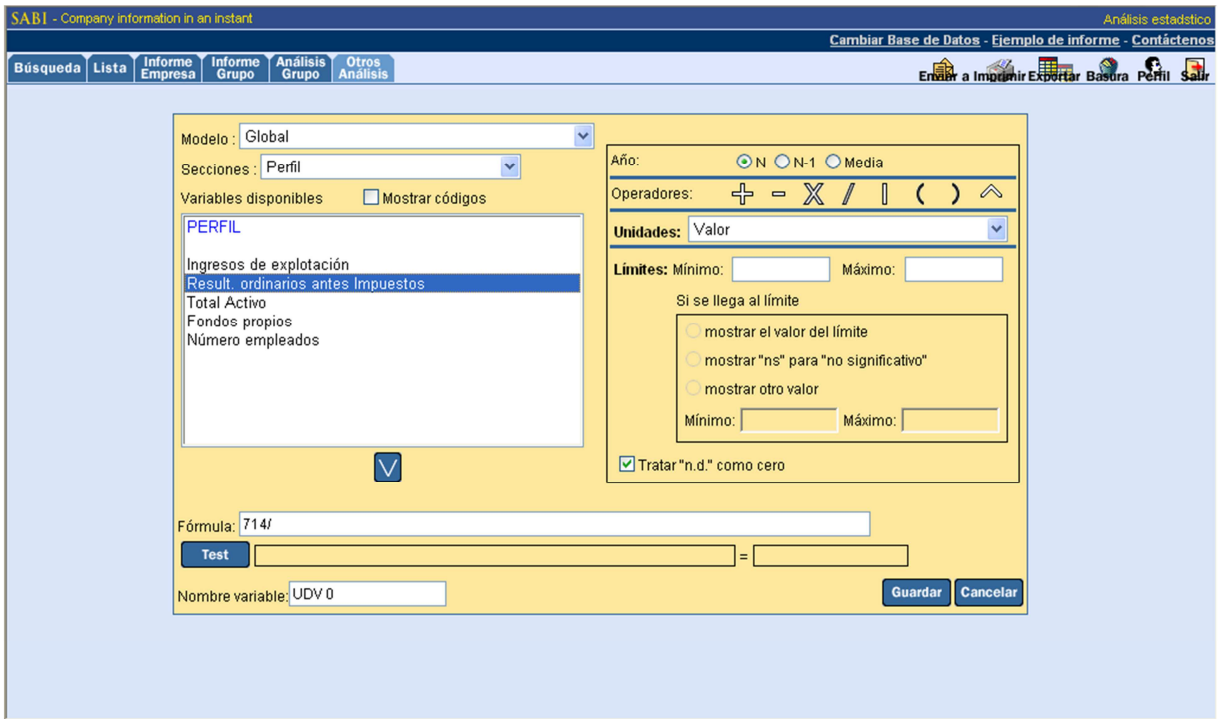


Ilustración 19: SABI. Posibilidad de creación de ratios personalizados con su formulación a partir de las distintas variables.



5.2. Aprendizaje computacional

La Inteligencia Artificial es una rama de las ciencias computacionales que trata de modelar el comportamiento humano por medio de la creación de sistemas que sean capaces de imitar la comprensión humana y que también sean capaces de aprender y reconocer.

Dentro de la Inteligencia Artificial, las distintas técnicas tienen por objetivo desarrollar aplicaciones para dar una solución a problemas que presentan un cierto de grado de complejidad, en los que a veces la solución óptima no es factible de encontrar, e incluso desconocida, por lo que estas aplicaciones nos ofrecen soluciones parciales que suelen dar buenos resultados.

Para construir dichas aplicaciones, en la programación se deberá incluir conocimiento del dominio del problema para facilitar su resolución, logrando con esto la reducción del costo computacional que implica el uso de técnicas convencionales.

Con frecuencia estas aplicaciones suelen ser programadas para dar solución a un problema en específico, sin embargo al momento que se quiere generalizar el uso de estas para problemas similares, se necesita volver a programar y agregar conocimiento nuevo para resolver estos problemas.

Vemos entonces que una de sus principales limitaciones radica en que no podrán resolver problemas para los que no hayan sido programadas. Es aquí donde el conocimiento que vayamos integrando aumenta el uso que le podemos dar a estas aplicaciones.

Una aplicación que sea capaz de adaptarse y poder integrar nuevo conocimiento, a medida que se usa, de manera automática se considera más cercana a ser realmente una aplicación inteligente, ya que podrá resolver nuevos problemas que nosotros le proporcionemos sin esperar una reprogramación y un conocimiento nuevo que le ayude a atacar estos problemas. El área de investigación llamada Aprendizaje Computacional, dentro de la Inteligencia Artificial, se encarga de desarrollar técnicas para construir aplicaciones en las que se tenga una adaptación de conocimiento de manera automática.

La idea del Aprendizaje Computacional es buscar métodos que logren aumentar las capacidades de las aplicaciones habituales de manera que puedan ser más flexibles y eficaces. Uso del aprendizaje computacional para la construcción de programas de inteligencia artificial:

- Tareas difíciles de programar: Es muy frecuente que nos encontremos con tareas excesivamente complejas en las que construir un programa para resolverlas resulta muy difícil. Pongamos el ejemplo en el que se pretende desarrollar un sistema de visión que sea capaz de reconocer un conjunto de imágenes, sería muy complicado desarrollar un programa a mano que resuelva este reconocimiento. El aprendizaje computacional nos permite construir un modelo de clasificación a partir de un conjunto de ejemplos, para realizar la tarea del reconocimiento y de esta manera hacer la resolución del problema más sencilla.
- Aplicaciones auto adaptables: Hay ciertas aplicaciones que tienen un mejor rendimiento si son capaces de adaptarse a las circunstancias e ir aprendiendo a lo largo de su ejecución. Podemos tener por ejemplo aplicaciones de interfaces que sean adaptables al uso continuo del usuario y poder ofrecer un mejor servicio.
- Minería de Datos/Descubrimiento de conocimiento: El aprendizaje puede servir para ayudar a analizar información, extrayendo de manera automática conocimiento a partir de conjuntos de ejemplos y descubriendo patrones complejos.

El aprendizaje computacional se divide en cuatro tipos de aprendizaje:

- Aprendizaje Inductivo: En este tipo de aprendizaje se crean modelos de conceptos o descriptores que posean características comunes de los ejemplos de entrenamiento. Aprendizaje analítico o deductivo: Es aplicada la deducción para obtener descriptores de un ejemplo, de un concepto y su descripción.

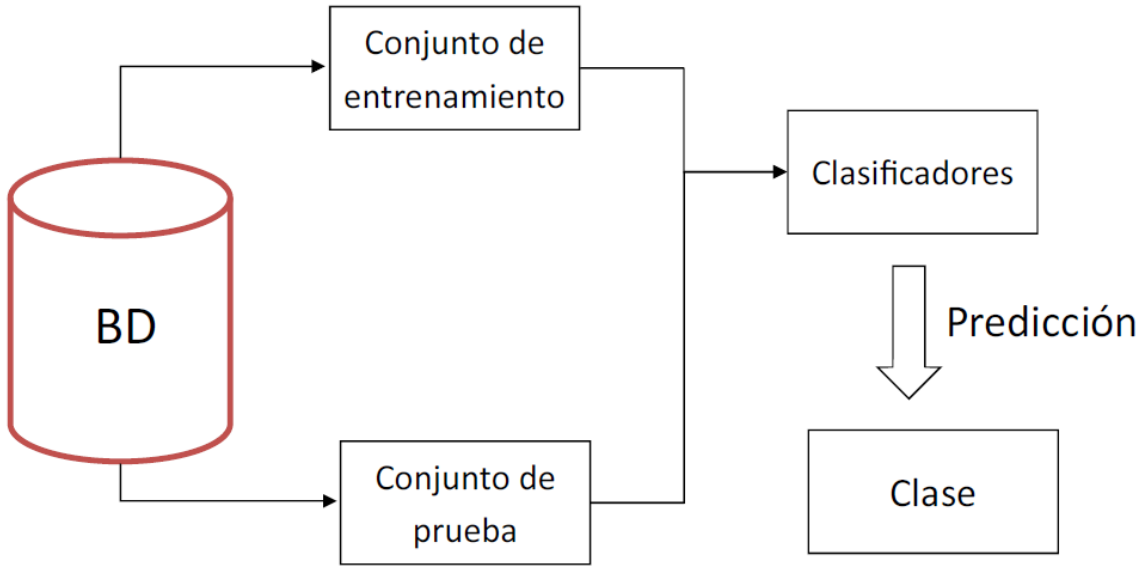
- Aprendizaje genético: Son aplicados algoritmos inspirados en la Teoría de la evolución para encontrar descriptores generales de los ejemplos.
- Aprendizaje conexionista: Se desarrollan descriptores generales mediante el uso de las capacidades de adaptación de redes neuronales artificiales. Una red neuronal está compuesta de elementos simples interconectados que poseen algún estado. Tras un proceso de entrenamiento, el estado en el que quedan las neuronas de la red representa el concepto aprendido.

En este estudio nos basaremos en el aprendizaje inductivo, en donde de distingue el aprendizaje supervisado y el no supervisado:

- Aprendizaje supervisado: En este tipo de aprendizaje se tiene un conjunto de ejemplos, los cuales tienen asociados un atributo llamado la clase de cada ejemplo, tendremos entonces algún mecanismo de entrenamiento en base a este conjunto, para permitirnos distinguir las clases sobre ejemplos de prueba que proporcionemos.
- Aprendizaje no supervisado: En este tipo de aprendizaje solo se cuenta con un conjunto de ejemplos que no tiene asociado ninguna clase, por lo que se debe de encontrar una manera de agruparlos. Para los métodos de aprendizaje no supervisado se utiliza el concepto de similaridad/disimilaridad de los ejemplos, para poder construir grupos en los que ejemplos similares estén justos en un grupo y separados de otros ejemplos menos similares.

En la Ilustración 20 se muestra un esquema básico de un clasificador en el aprendizaje computacional, que muestra como de una BD se seleccionan dos conjuntos, uno de entrenamiento para aplicar una fase de entrenamiento a cada clasificador utilizado, para que de esta forma se pueda pasar a clasificar el conjunto de prueba del cual desconocemos la clase de cada ejemplo de prueba.

Ilustración 20: Esquema de un clasificador en aprendizaje computacional



.

A continuación se describen la forma de clasificar a un conjunto de ejemplos mediante construcción de Árboles de Decisión con el algoritmo ID3 (Iterative Dichotomiser 3) y C4.5.

La construcción de árboles de decisión nos permite representar conocimiento a partir de un conjunto de ejemplos, con esto podemos realizar una generalización de ellos y hallar una forma de clasificar los ejemplos dados. Se analizará un ejemplo práctico de cómo se construyen estos árboles y se finalizará con algunas ventajas y desventajas de este método de clasificación.

Un árbol de decisión es una forma gráfica y analítica para poder llevar a cabo la clasificación de los datos utilizados, mediante diferentes caminos posibles. Cada uno de los nodos del árbol representa los diferentes atributos presentes en los datos, las ramificaciones del árbol representan los caminos posibles a seguir, para predecir la clase de un nuevo ejemplo, en donde los nodos terminales u hojas establecen la clase a la que pertenece el ejemplo de prueba si se sigue por la ramificación en cuestión.

El lenguaje de descripción de los árboles de decisión corresponde a las fórmulas en FND (Forma Normal Disyuntiva). Consideramos el caso en el que tenemos 3 atributos,

el atributo A, B y C, cada uno de ellos con dos valores, $x_i \; y \; \neg x_i \;\; \forall i \equiv 1,2,3$ respectivamente, con esto se pueden construir 2^n combinaciones en FNC (Forma Normal Conjuntiva). Un ejemplo de estas se muestra a continuación:

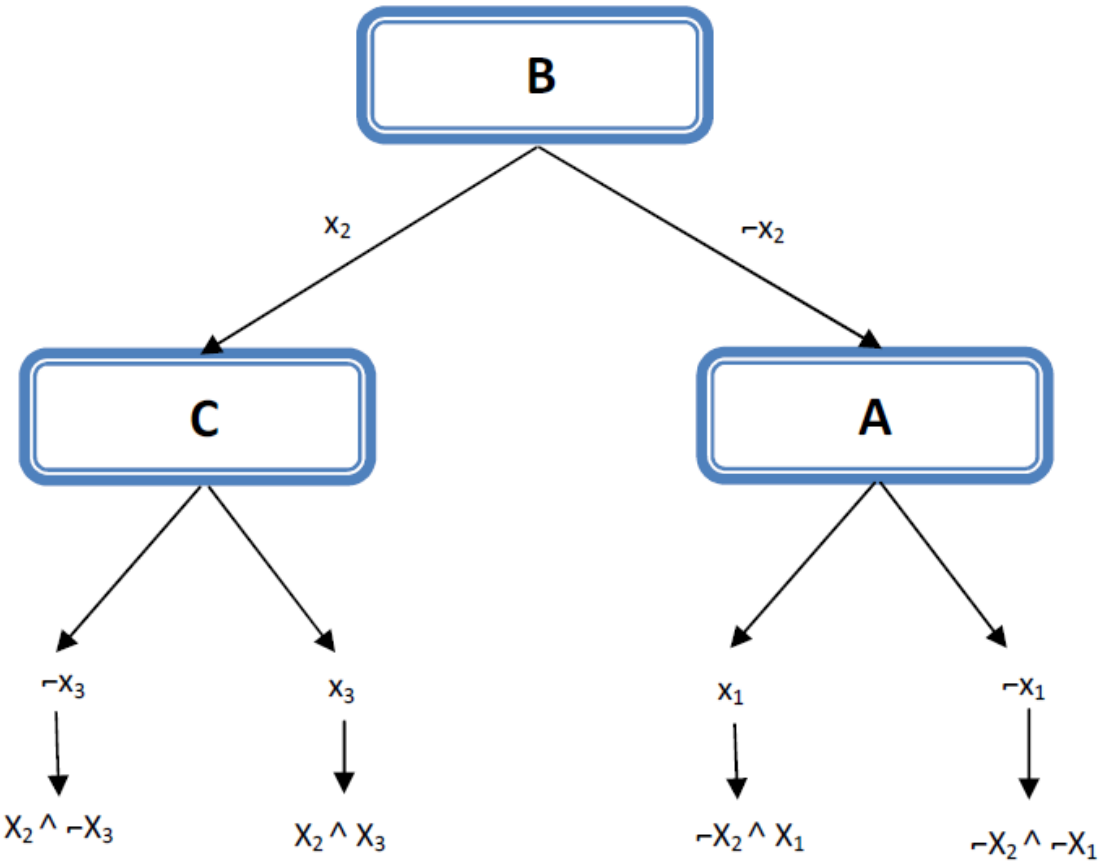
$$(x_1 \cap x_2 \cap \neg x_3)$$

Cada una de las combinaciones en FNC describe una parte del árbol que estemos formando, por lo que tendríamos para el árbol disyuntivas de la siguiente forma:

$$(x_2 \cap \neg x_3) \cup (x_2 \cap x_3) \cup (\neg x_2 \cap x_1) \cup (\neg x_2 \cap \neg x_1)$$

Las cuales son descriptores del árbol de inducción construido, entonces podríamos formar $O(2^{2n})$ descripciones posibles en FND. El árbol correspondiente a los descriptores anteriores se muestra en la Ilustración 12.

Ilustración 21: Ejemplo gráfico Árbol de decisión



Dado que el orden de los árboles o descriptores es muy grande, evidentemente no es posible explorar todos los descriptores para ver cuál es el más adecuado, por lo que se utilizan técnicas de búsqueda heurística para encontrar una forma más fácil y rápida de hacerlo.

La mayoría de los algoritmos de construcción de árboles de decisión se basa en la estrategia de Ascenso a la Colina (Hill Climbing). Esta es una técnica utilizada en Inteligencia Artificial para encontrar los máximos o mínimos de una función mediante una búsqueda local. Estos algoritmos empiezan con un árbol vacío, después se va particionado en conjuntos de ejemplos, eligiendo en cada caso aquel atributo que mejor discrimina entre las clases, hasta que se completa el árbol.

Para saber qué atributo es el mejor se utiliza una función heurística, la elección que hagamos es irrevocable, por lo que debemos asegurar que esta sea la más cercana a la óptima. La principal ventaja de usar este tipo de estrategias es que el costo computacional es bastante reducido.

5.2.1. Algoritmo ID3 (Iterative Dichotomiser 3)

El algoritmo ID3 desarrollado por Quinlan en 1983 es considerado un algoritmo seminal, ya que de aquí se derivan muchos algoritmos para la construcción de árboles de decisión. Este algoritmo se basa en la teoría de la información, desarrollada en 1948 por Claude Elwood Shannon.

Su significado es “inducción mediante árboles de decisión”, capaz de tomar decisiones con gran precisión.

Es un sistema de aprendizaje supervisado que aplica la estrategia “divide y vencerás” para hacer la clasificación, implementando métodos y técnicas para la realización de procesos inteligentes, representando así el conocimiento y el aprendizaje, con el propósito de automatizar tareas.

El ID3 permite determinar el árbol de decisión mínimo, para un conjunto de objetos. Permite que la información se mantenga organizada y entendible para cualquier

persona, además haciendo uso de una secuencia de preguntas, donde cada pregunta es evaluada con el propósito de obtener la mejor respuesta posible.

La idea básica del ID3 es de determinar, para un conjunto de ejemplos dados, el atributo más importante, es decir, aquel que posea mayor poder discriminatorio para dicho conjunto; este atributo es usado para la clasificación de la lista de objetos, basados en los valores asociados por él mismo. Después de haber hecho la primera prueba de atributo, esta arrojará un resultado, el cual es en sí mismo un nuevo problema de aprendizaje de árbol de decisión con la diferencia de que contará con menos ejemplos y un atributo menos, por lo que cada atributo que se selecciona se descarta para la siguiente prueba.

Se utiliza la noción de entropía descrita en la teoría de la información, para ver que aleatoriedad presenta la distribución de un conjunto de ejemplos sobre las clases a las que pertenecen. Dentro de esta teoría de la información se estudia también cuales son los mecanismos de codificación de los mensajes y el costo asociado a su transmisión.

Sea $M = \{m_1, m_2, \dots\dots, m_n\}$ un conjunto de mensajes, en donde para cada uno de ellos se tiene una probabilidad $P(m_i)$, entonces podemos definir la cantidad de información I contenida en un mensaje de M como:

$$I(M) = \sum_{i=1}^{n} -P(m_i)\log_2\big(P(m_i)\big)$$

Se multiplica por menos, debido a que como manejamos probabilidades que están definidas entre [0, 1], el logaritmo de estas probabilidades nos da un número negativo, por lo que convertimos este valor en positivo para darle un sentido intuitivo.

Este valor nos dice que cantidad de bits de información son necesarios para codificar los diferentes mensajes de M. Dado un conjunto de mensajes, podemos obtener la mejor manera de codificarlos para que el coste de su transmisión sea mínimo. Además nos permite ir seleccionando al mejor atributo cuyo conocimiento aporte mayor información desde la perspectiva de clasificación, en cada uno de los niveles del árbol que se vaya construyendo. Podemos realizar una analogía con la codificación de

mensajes tomando a las clases como los mensajes y la proporción de ejemplos de cada clase como su probabilidad. El objetivo de los árboles de decisión es construir el árbol de tamaño mínimo, que nos permita distinguir los ejemplos de cada clase. Cada atributo se deberá evaluar para decidir si se le incluye en el árbol. Un atributo será mejor cuanto más permita discriminar entre las diferentes clases como se había comentado anteriormente.

Se parte de un árbol vacío y se va construyendo de manera recursiva, tomando en cada nodo aquel atributo que tiene el mayor grado de información, haciendo que sea menos la cantidad de información que falta por cubrir.

Cada vez que hagamos la elección de un atributo, debería hacer que los subconjuntos de ejemplos que genera el atributo sean mayoritariamente de una clase. Para medir esto necesitamos una medida de la cantidad de información que cubre un atributo (medida de Entropía, E). Lo primero que se debe formular es la cantidad de información que tiene cada clase la cual se puede hacer de la siguiente forma:

Dado un conjunto de ejemplos X clasificados en un conjunto de clases $C = \{c_1, c_2, \dots\dots, c_n\}$ siendo $|c_i|$ la cardinalidad de la clase c_i y $|X|$ el número total de ejemplos, la cantidad de información vendrá expresada de la siguiente forma:

$$I(X, C) = -\sum_{c_i \in C} \frac{|c_i|}{|X|} \log_2 \left(\frac{|c_i|}{|X|} \right)$$

Teniendo la cantidad de información se puede formular la Entropía de la siguiente forma:

Para cada uno de los atributos A_i; siendo $\{v_{i1}, \dots\dots, v_{in}\}$ el conjunto de posibles valores del atributo A_i y $\left|\left[A_i(C) = v_{ij}\right]\right|$ el número de ejemplos que tienen el valor v_{ij} en su atributo A_i, la función de entropía es expresada de la siguiente forma:

$$E(X, C, A_i) = -\sum_{v_{ij} \in A_i} \frac{\left|\left[A_i(C) = v_{ij}\right]\right|}{|X|} I\left(\left[A_i(C) = v_{ij}\right], C\right)$$

Con el resultado de estas medidas se define la función de ganancia de información de la siguiente forma:

$$G(|X|, C, A_i) = I(X, C) - E(X, C, A_i)$$

Es por ello que se toma el atributo que maximice este valor para ir expandiendo el árbol de inducción. El pseudocódigo del algoritmo ID3 para la construcción de un árbol de decisión es el siguiente:

func $ID3(X, C, A)$ $\equiv$ (X: Ejemplos, C: Clasificación, A: Atributos)

  si todos los ejemplos son de la misma clase

    entonces $ID3 \leftarrow$ hoja con la clase.

  otro

    Calcular la función de cantidad de información de los ejemplos (I)

    para cada atributo en A

      Calcular la función de entropía (E) y la ganancia de información (G)

      Escoger el atributo que maximiza (G) (sea a)

      Eliminar a de la lista de atributos (A)

    termina

    para cada partición generada por los valores v_i del atributo a

      $Arbol_i \leftarrow ID3$(ejemplos de X con $a = v_i$,

      Clasificación de los ejemplos, Atributos restantes)

      Generar árbol con $a = v_i$y $Arbol_i$

    termina

    $ID3 \leftarrow$ la unión de todos los árboles

A continuación implementamos un ejemplo mediante el Algoritmo ID3 de forma manual.

Se han escogido de los estudios analizados los ratios más importantes y se han establecido una serie de ejemplos aleatorios de empresas ficticias que se muestran en la tabla 21 a partir de las que se obtiene la Ilustración 13.

Tabla 19: Ejemplo Implementación Algoritmo ID3 Tabla de empresas y ratios

	Atributo				
Empresa	Rentabilidad	Cash Flow	Endudamiento	Fondos Propios	Solvencia
E1	Alta	Alto	Alto	Bajo	Insolvente
E2	Alta	Alto	Alto	Alto	Solvente
E3	Media	Alto	Alto	Bajo	Insolvente
E4	Baja	Medio	Alto	Bajo	Insolvente
E5	Baja	Bajo	Bajo	Bajo	Insolvente
E6	Baja	Bajo	Bajo	Alto	Solvente
E7	Media	Bajo	Bajo	Alto	Solvente
E8	Alta	Medio	Alto	Bajo	Insolvente
E9	Alta	Bajo	Bajo	Bajo	Solvente
E10	Baja	Medio	Bajo	Bajo	Insolvente
E11	Alta	Medio	Bajo	Alto	Solvente
E12	Media	Medio	Alto	Alto	Solvente
E13	Media	Alto	Bajo	Bajo	Solvente
E14	Baja	Medio	Alto	Alto	Insolvente

Tabla 20: Cálculo ejemplo algoritmo ID3 (1 de 6)

Información de las clases								
	Insolvente	.=7/14					**Variable**	**Ganancia de Información (Variable)**
Clase	**Solvente**	.=7/14					Rentabilidad	0,163606718
	I(S)	-(7/14)*LOG(7/14;2)-(7/14)*LOG(7/14;2)=			1		Cash Flow	0,088936607
							Endudamiento	0,136879431
Rentabilidad	**(Alta, media, Baja)**						Fondos Propios	0,199883025
PyA=Alta		Insolvente	.=2/5					
	Alta	Solvente	.=3/5					F23-((5/14)*F28+(4/14)*F32+(5/14)*F36)=
	I(S,Rentabilidad)	-(2/5)*LOG(2/5;2)-(3/5)*LOG(3/5;2)=			0,97095059			F23-((4/14)*F41+(6/14)*F45+(4/14)*F49)=
								=F23-((7/14)*F54+(7/14)*F58)=
PyA=Media		Insolvente	.=1/4					F23-((6/14)*F63+(9/14)*F67)=
	Media	Solvente	.=3/4					
	I(S,Rentabilidad)	-(1/4)*LOG(1/4;2)-(3/4)*LOG(3/4;2)=			0,81127812			
PyA=Baja		Insolvente	.=4/5					
	Baja	Solvente	.=1/5					
	I(S,Rentabilidad)	-(4/5)*LOG(4/5;2)-(1/5)*LOG(1/5;2)=			0,72192809			
Cash Flow	**(Alto, medio, Bajo)**							
PyA=Alto		Insolvente	.=2/4					
	Alto	Solvente	.=2/4					
	I(S,Rentabilidad)	-(2/4)*LOG(2/4;2)-(2/4)*LOG(2/4;2)=			1			
PyA=Medio		Insolvente	.=4/6					
	Medio	Solvente	.=2/6					
	I(S,Rentabilidad)	-(4/6)*LOG(4/6;2)-(2/6)*LOG(2/6;2)=			0,91829583			
PyA=Bajo		Insolvente	.=1/4					
	Bajo	Solvente	.=3/4					
	I(S,Rentabilidad)	-(1/4)*LOG(1/4;2)-(3/4)*LOG(3/4;2)=			0,81127812			
Endeudamiento	**(Alto,Bajo)**							
PyA=Alto		Insolvente	.=5/7					
	Alto	Solvente	.=2/7					
	I(S,Rentabilidad)	-(5/7)*LOG(5/7;2)-(2/7)*LOG(2/7;2)=			0,86312057			
PyA=Bajo		Insolvente	.=2/7					
	Bajo	Solvente	.=5/7					
	I(S,Rentabilidad)	-(2/7)*LOG(2/7;2)-(5/7)*LOG(5/7;2)=			0,86312057			
Fondos Propios	**(Si/No)**							
PyA=Alto		Insolvente	.=1/6					
	Alto	Solvente	.=5/6					
	I(S,Rentabilidad)	-(1/6)*LOG(1/6;2)-(5/6)*LOG(5/6;2)=			0,65002242			
PyA=Bajo		Insolvente	.=6/8					
	Bajo	Solvente	.=2/8					
	I(S,Rentabilidad)	-(6/8)*LOG(6/8;2)-(2/8)*LOG(2/8;2)=			0,81127812			

Tabla 21: Cálculo ejemplo algoritmo ID3 (2 de 6)

	Atributo				
Empresa	**Rentabilidad**	**Cash Flow**	**Endudamiento**	**Fondos Propios**	**Solvencia**
E1	Alta	Alto	Alto	Bajo	Insolvente
E2	Alta	Alto	Alto	Alto	Solvente
E3	Media	Alto	Alto	Bajo	Insolvente
E4	Baja	Medio	Alto	Bajo	Insolvente
E5	Baja	Bajo	Bajo	Bajo	Insolvente
E6	Baja	Bajo	Bajo	Alto	Solvente
E7	Media	Bajo	Bajo	Alto	Solvente
E8	Alta	Medio	Alto	Bajo	Insolvente
E9	Alta	Bajo	Bajo	Bajo	Solvente
E10	Baja	Medio	Bajo	Bajo	Insolvente
E11	Alta	Medio	Bajo	Alto	Solvente
E12	Media	Medio	Alto	Alto	Solvente
E13	Media	Alto	Bajo	Bajo	Solvente
E14	Baja	Medio	Alto	Alto	Insolvente

Información de las clases		FP Bajo (Elimino Alto)		
	Insolvente	.=6/8		
Clase	**Solvente**	.=2/8		
	I(S)	-(6/8)*LOG(6/8;2)-(2/8)*LOG(2/8;2)=		0,81127812
Rentabilidad	**(Alta, media, Baja)**			
PyA=Alta		Insolvente	.=2/3	
	Alta	Solvente	.=1/3	
	I(S,Rentabilidad)	-(2/3)*LOG(2/3;2)-(1/3)*LOG(1/3;2)=		0,91829583
PyA=Media		Insolvente	.=1/2	
	Media	Solvente	.=1/2	
	I(S,Rentabilidad)	-(1/2)*LOG(1/2;2)-(1/2)*LOG(1/2;2)=		1
PyA=Baja		Insolvente	.=3/3	
	Baja	Solvente	.=0/3	
	I(S,Rentabilidad)	-(3/3)*LOG(3/3;2)-(0/3)*LOG(0/3;2)=		0
Cash Flow	**(Alto, medio, Bajo)**			
PyA=Alto		Insolvente	.=2/3	
	Alto	Solvente	.=1/3	
	I(S,Rentabilidad)	-(2/3)*LOG(2/3;2)-(1/3)*LOG(1/3;2)=		0,91829583
PyA=Medio		Insolvente	.=3/3	
	Medio	Solvente	.=0/3	
	I(S,Rentabilidad)	-(3/3)*LOG(3/3;2)-(0/3)*LOG(0/3;2)=		0
PyA=Bajo		Insolvente	.=1/2	
	Bajo	Solvente	.=1/2	
	I(S,Rentabilidad)	-(1/2)*LOG(1/2;2)-(1/2)*LOG(1/2;2)=		1
Endeudamiento	**(Alto,Bajo)**			
PyA=Alto		Insolvente	.=4/4	
	Alto	Solvente	.=0/4	
	I(S,Rentabilidad)	-(4/4)*LOG(4/4;2)-(0/4)*LOG(0/4;2)=		0
PyA=Bajo		Insolvente	.=2/4	
	Bajo	Solvente	.=2/4	
	I(S,Rentabilidad)	-(2/4)*LOG(2/4;2)-(2/4)*LOG(2/4;2)=		1

Variable	Ganancia de Información (Variable)
Rentabilidad	0,216917187
Cash Flow	0,216917187
Endudamiento	0,311278124

Tabla 22: Cálculo ejemplo algoritmo ID3 (3 de 6)

	Atributo				
Empresa	Rentabilidad	Cash Flow	Endudamien	Fondos Pro	Solvencia
E1	Alta	Alto	Alto	Bajo	Insolvente
E2	Alta	Alto	Alto	Alto	Solvente
E3	Media	Alto	Alto	Bajo	Insolvente
E4	Baja	Medio	Alto	Bajo	Insolvente
E5	Baja	Bajo	Bajo	Bajo	Insolvente
E6	Baja	Bajo	Bajo	Alto	Solvente
E7	Media	Bajo	Bajo	Alto	Solvente
E8	Alta	Medio	Alto	Bajo	Insolvente
E9	Alta	Bajo	Bajo	Bajo	Solvente
E10	Baja	Medio	Bajo	Bajo	Insolvente
E11	Alta	Medio	Bajo	Alto	Solvente
E12	Media	Medio	Alto	Alto	Solvente
E13	Media	Alto	Bajo	Bajo	Solvente
E14	Baja	Medio	Alto	Alto	Insolvente

Información de las clases			FP Alto (Elimino Bajo)	
	Insolvente	.=1/6		
Clase	**Solvente**	.=5/6		
	I(S)	-(1/6)*LOG(1/6;2)-(5/6)*LOG(5/6;2)=		0,65002242
Rentabilidad	**(Alta, media, Baja)**			
PyA=Alta		Insolvente	.=0/2	
	Alta	Solvente	.=2/2	
	I(S,Rentabilidad	-(0/2)*LOG(0/2;2)-(2/2)*LOG(2/2;2)=		0
PyA=Media		Insolvente	.=0/2	
	Media	Solvente	.=2/2	
	I(S,Rentabilidad	-(0/2)*LOG(0/2;2)-(2/2)*LOG(2/2;2)=		0
PyA=Baja		Insolvente	.=1/2	
	Baja	Solvente	.=1/2	
	I(S,Rentabilidad	-(1/2)*LOG(1/2;2)-(1/2)*LOG(1/2;2)=		1
Cash Flow	**(Alto, medio, Bajo)**			
PyA=Alto		Insolvente	.=0/1	
	Alto	Solvente	.=1/1	
	I(S,Rentabilidad	-(1/1)*LOG(1/1;2)=		0
PyA=Medio		Insolvente	.=1/3	
	Medio	Solvente	.=2/3	
	I(S,Rentabilidad	-(1/3)*LOG(1/3;2)-(2/3)*LOG(2/3;2)=		0,91829583
PyA=Bajo		Insolvente	.=0/2	
	Bajo	Solvente	.=2/2	
	I(S,Rentabilidad	-(2/2)*LOG(2/2;2)=		0
Endeudamie	**(Alto,Bajo)**			
PyA=Alto		Insolvente	.=1/3	
	Alto	Solvente	.=2/3	
	I(S,Rentabilidad	-(1/3)*LOG(1/3;2)-(2/3)*LOG(2/3;2)=		0,91829583
PyA=Bajo		Insolvente	.=0/3	
	Bajo	Solvente	.3/3	
	I(S,Rentabilidad	-(3/3)*LOG(3/3;2)=		0

Variable	Ganancia de Información (Variable)
Rentabilidad	0,316689088
Cash Flow	0,190874505
Endudamient	0,190874505

Tabla 23: Cálculo ejemplo algoritmo ID3 (4 de 6)

	Atributo				
Empresa	**Rentabilidad**	**Cash Flow**	**Endudamiento**	**Fondos Propios**	**Solvencia**
E1	Alta	Alto	Alto	Bajo	Insolvente
E2	Alta	Alto	Alto	Alto	Solvente
E3	Media	Alto	Alto	Bajo	Insolvente
E4	Baja	Medio	Alto	Bajo	Insolvente
E5	Baja	Bajo	Bajo	Bajo	Insolvente
E6	Baja	Bajo	Bajo	Alto	Solvente
E7	Media	Bajo	Bajo	Alto	Solvente
E8	Alta	Medio	Alto	Bajo	Insolvente
E9	Alta	Bajo	Bajo	Bajo	Solvente
E10	Baja	Medio	Bajo	Bajo	Insolvente
E11	Alta	Medio	Bajo	Alto	Solvente
E12	Media	Medio	Alto	Alto	Solvente
E13	Media	Alto	Bajo	Bajo	Solvente
E14	Baja	Medio	Alto	Alto	Insolvente

Información de las clases		Endeudamiento Bajo (Elimino Alto)				
	Insolvente	.=2/7				
Clase	**Solvente**	.=5/7				
	I(S)	-(2/7)*LOG(2/7;2)-(5/7)*LOG(5/7;2)=			0,86312057	
Rentabilidad	**(Alta, media, Baja)**					
PyA=Alta		Insolvente	.=0/2			
	Alta	Solvente	.=2/2			
	I(S,Rentabilidad)	-(2/2)*LOG(2/2;2)=			0	
PyA=Media		Insolvente	.=0/2			
	Media	Solvente	.=2/2			
	I(S,Rentabilidad)	-(2/2)*LOG(2/2;2)=			0	
PyA=Baja		Insolvente	.=2/3			
	Baja	Solvente	.=1/3			
	I(S,Rentabilidad)	-(2/3)*LOG(2/3;2)-(1/3)*LOG(1/3;2)=			0,91829583	
Cash Flow	**(Alto, medio, Bajo)**					
PyA=Alto		Insolvente	.=0/1			
	Alto	Solvente	.=1/1			
	I(S,Rentabilidad)	-(1/1)*LOG(1/1;2)=			0	
PyA=Medio		Insolvente	.=1/2			
	Medio	Solvente	.=1/2			
	I(S,Rentabilidad)	-(1/2)*LOG(1/2;2)-(1/2)*LOG(1/2;2)=			1	
PyA=Bajo		Insolvente	.=1/4			
	Bajo	Solvente	.=3/4			
	I(S,Rentabilidad)	-(1/4)*LOG(1/4;2)-(3/4)*LOG(3/4;2)=			0,81127812	

Variable	Ganancia de Información (Variable)
Rentabilidad	0,469565211
Cash Flow	0,113818783

Tabla 24: Cálculo ejemplo algoritmo ID3 (5 de 6)

	Atributo				
Empresa	Rentabilidad	Cash Flow	Endudamiento	Fondos Pro	Solvencia
E1	Alta	Alto	Alto	Bajo	Insolvente
E2	Alta	Alto	Alto	Alto	Solvente
E3	Media	Alto	Alto	Bajo	Insolvente
E4	Baja	Medio	Alto	Bajo	Insolvente
E5	Baja	Bajo	Bajo	Bajo	Insolvente
E6	Baja	Bajo	Bajo	Alto	Solvente
E7	Media	Bajo	Bajo	Alto	Solvente
E8	Alta	Medio	Alto	Bajo	Insolvente
E9	Alta	Bajo	Bajo	Bajo	Solvente
E10	Baja	Medio	Bajo	Bajo	Insolvente
E11	Alta	Medio	Bajo	Alto	Solvente
E12	Media	Medio	Alto	Alto	Solvente
E13	Media	Alto	Bajo	Bajo	Solvente
E14	Baja	Medio	Alto	Alto	Insolvente

Información de las clase		Endeudamiento Alto (Elimino Bajo)				
	Insolvente	.=5/7				
Clase	**Solvente**	.=2/7				
	I(S)	-(2/7)*LOG(2/7;2)-(5/7)*LOG(5/7;2)=			0,86312057	
Rentabilidad	**(Alta, media, Baja)**					
PyA=Alta		Insolvente	.=2/3			
	Alta	Solvente	.=1/3			
	I(S,Rentabilidad	-(2/3)*LOG(2/3;2)-(1/3)*LOG(1/3;2)=			0,91829583	
PyA=Media		Insolvente	.=1/2			
	Media	Solvente	.=1/2			
	I(S,Rentabilidad	-(1/2)*LOG(1/2;2)-(1/2)*LOG(1/2;2)=			1	
PyA=Baja		Insolvente	.=2/2			
	Baja	Solvente	.=0/2			
	I(S,Rentabilidad	-(2/2)*LOG(2/2;2)=			0	
Cash Flow	**(Alto, medio, Bajo)**					
PyA=Alto		Insolvente	.=1/3			
	Alto	Solvente	.=2/3			
	I(S,Rentabilidad	-(1/3)*LOG(1/3;2)-(2/3)*LOG(2/3;2)=			0,91829583	
PyA=Medio		Insolvente	.=3/4			
	Medio	Solvente	.=1/4			
	I(S,Rentabilidad	-(3/4)*LOG(3/4;2)-(1/4)*LOG(1/4;2)=			0,81127812	
PyA=Bajo		Insolvente	.=0			
	Bajo	Solvente	.=0			
	I(S,Rentabilidad	0			0	

Variable	Ganancia de Información (Variable)
Rentabilidad	0,183850925
Cash Flow	0,005977711

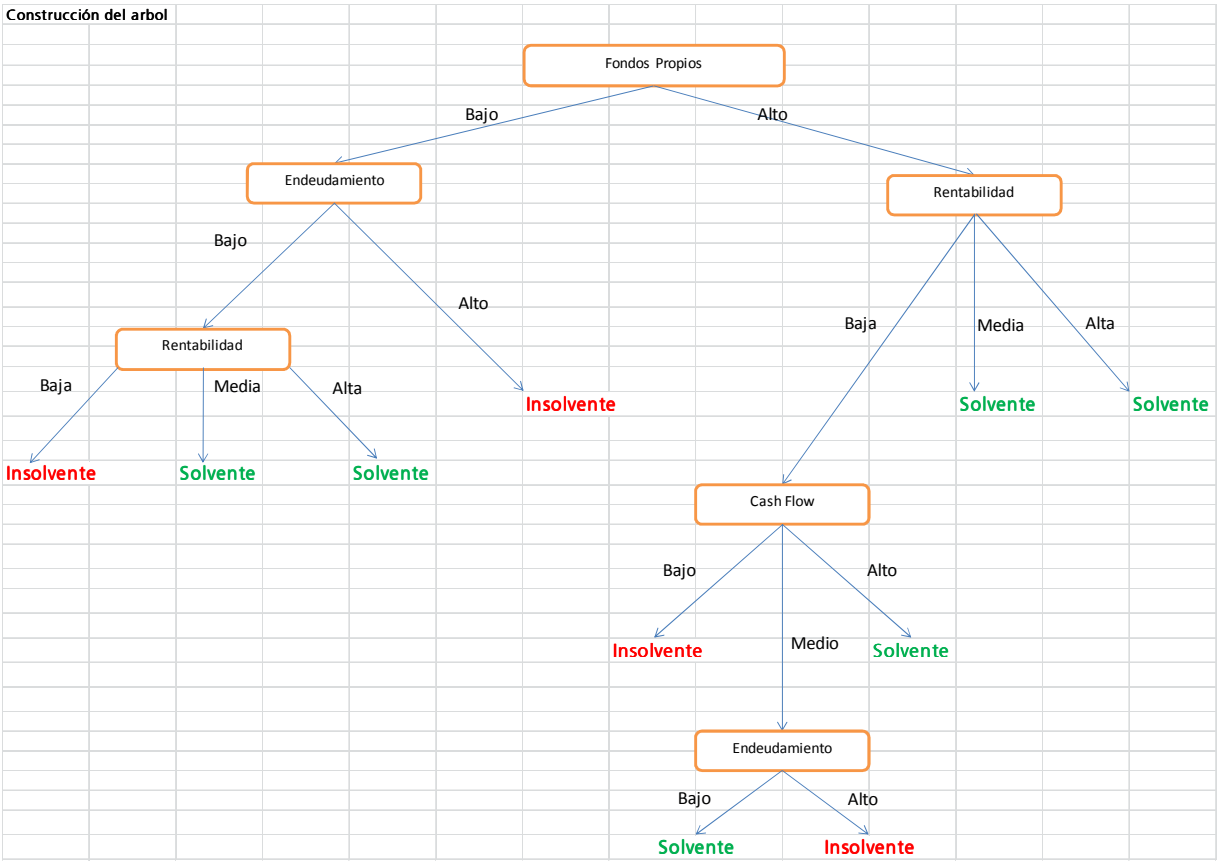
Tabla 25: Cálculo ejemplo algoritmo ID3 (6 de 6)

	Atributo				
Empresa	**Rentabilidad**	**Cash Flow**	**Endudamiento**	**Fondos Propios**	**Solvencia**
E1	Alta	Alto	Alto	Bajo	Insolvente
E2	Alta	Alto	Alto	Alto	Solvente
E3	Media	Alto	Alto	Bajo	Insolvente
E4	Baja	Medio	Alto	Bajo	Insolvente
E5	Baja	Bajo	Bajo	Bajo	Insolvente
E6	Baja	Bajo	Bajo	Alto	Solvente
E7	Media	Bajo	Bajo	Alto	Solvente
E8	Alta	Medio	Alto	Bajo	Insolvente
E9	Alta	Bajo	Bajo	Bajo	Solvente
E10	Baja	Medio	Bajo	Bajo	Insolvente
E11	Alta	Medio	Bajo	Alto	Solvente
E12	Media	Medio	Alto	Alto	Solvente
E13	Media	Alto	Bajo	Bajo	Solvente
E14	Baja	Medio	Alto	Alto	Insolvente

Información de las clases		Rentabilidad Baja (Elimino Alta y Media)			
	Insolvente	.=4/5			
Clase	**Solvente**	.=1/5			
	I(S)	-(4/5)*LOG(4/5;2)-(1/5)*LOG(1/5;2)=		0,72192809	
Cash Flow	**(Alto, medio, Bajo)**				
PyA=Alto		Insolvente	.=0		
	Alto	Solvente	.=0		
	I(S,Rentabilidad)	0		0	
PyA=Medio		Insolvente	.=3/3		
	Medio	Solvente	.=0/3		
	I(S,Rentabilidad)	-(3/3)*LOG(3/3;2)-(0/3)*LOG(0/3;2)=		0	
PyA=Bajo		Insolvente	.=1/2		
	Bajo	Solvente	.=1/2		
	I(S,Rentabilidad)	-(1/2)*LOG(1/2;2)-(1/2)*LOG(1/2;2)=		1	
Endeudamiento	**(Alto,Bajo)**				
PyA=Alto		Insolvente	.=2/2		
	Alto	Solvente	.=0/2		
	I(S,Rentabilidad)	-(2/2)*LOG(2/2;2)-(0/2)*LOG(0/2;2)=		0	
PyA=Bajo		Insolvente	.=2/3		
	Bajo	Solvente	.=1/3		
	I(S,Rentabilidad)	-(2/3)*LOG(2/3;2)-(1/3)*LOG(1/3;2)=		0,91829583	

Variable	Ganancia de Información (Variable)
Cash Flow	0,321928095
Endudamiento	0,170950594

Ilustración 22: Árbol de decisión resultante



5.3. Lenguaje de programación

El presente trabajo se desarrolla bajo el lenguaje de programación R, que incluye el paquete *rpart* que permite construir dichos modelos.

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. El paquete estadístico R es uno de los más flexibles, potentes y profesionales que existen actualmente para realizar tareas estadísticas de todo tipo, desde las más elementales, hasta las más avanzadas. En particular, está desarrollado y mantenido por algunos de los más prestigiosos estadísticos actuales.

Se trata de un proyecto de software libre, resultado de la implementación GNU del lenguaje S. R y S-Plus son, dos de los lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o gráfico.

R se distribuye bajo la licencia GNU GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.

R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas.

Al igual que S, se trata de un lenguaje de programación, lo que permite que los usuarios lo extiendan definiendo sus propias funciones. De hecho, gran parte de las funciones de R están escritas en el mismo R, aunque para algoritmos computacionalmente exigentes es posible desarrollar bibliotecas en C, C++ o Fortran que se cargan dinámicamente. También se pueden manipular los objetos de R directamente desde código desarrollado en C. R también puede extenderse a través de paquetes desarrollados por su comunidad de usuarios.

Además, R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python.

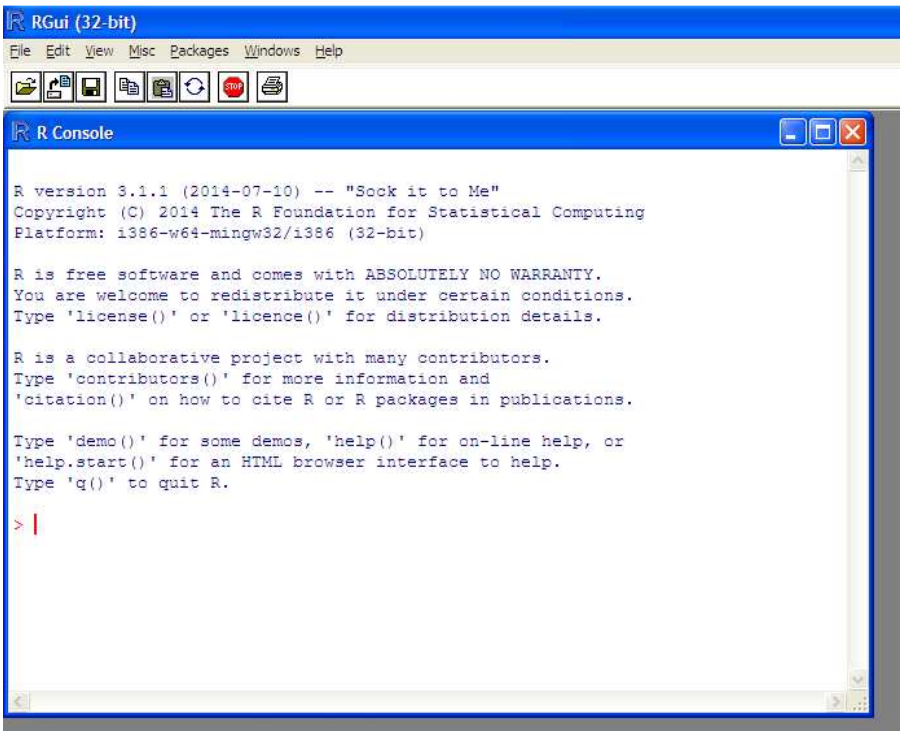
Otra de las características de R es su capacidad gráfica, que permite generar gráficos con alta calidad. R posee su propio formato para la documentación basado en *LaTeX.*

R también puede usarse como herramienta de cálculo numérico, campo en el que puede ser tan eficaz como otras herramientas específicas tales como *GNU Octave* y su equivalente comercial, MATLAB. Se ha desarrollado una interfaz, *RWeka6* para interactuar con *Weka* que permite leer y escribir ficheros en el formato *arff* y enriquecer R con los algoritmos de minería de datos de dicha plataforma.

R forma parte de un proyecto colaborativo y abierto. Sus usuarios pueden publicar paquetes que extienden su configuración básica.

Para facilitar el desarrollo de nuevos paquetes, se ha puesto a servicio de la comunidad una forja de desarrollo que facilita las tareas relativas a dicho proceso.

Ilustración 23: Pantalla principal del programa R



Los usuarios de R disponen de una serie de algoritmos estándar para generar y manipular árboles de decisión. Los más habituales están contenidos en alguno de los siguientes paquetes:

- Rpart
- RWeka
- Tree
- mvpart
- partykit

Cada uno de ellos tiene un interfaz distinto y operaciones como las de realizar predicciones, dibujar los árboles, etc. exigen conocer funciones específicas. (Éste es, de hecho, un problema genérico de R derivado de su naturaleza cooperativa).

Ilustración 24: Paquetes de R

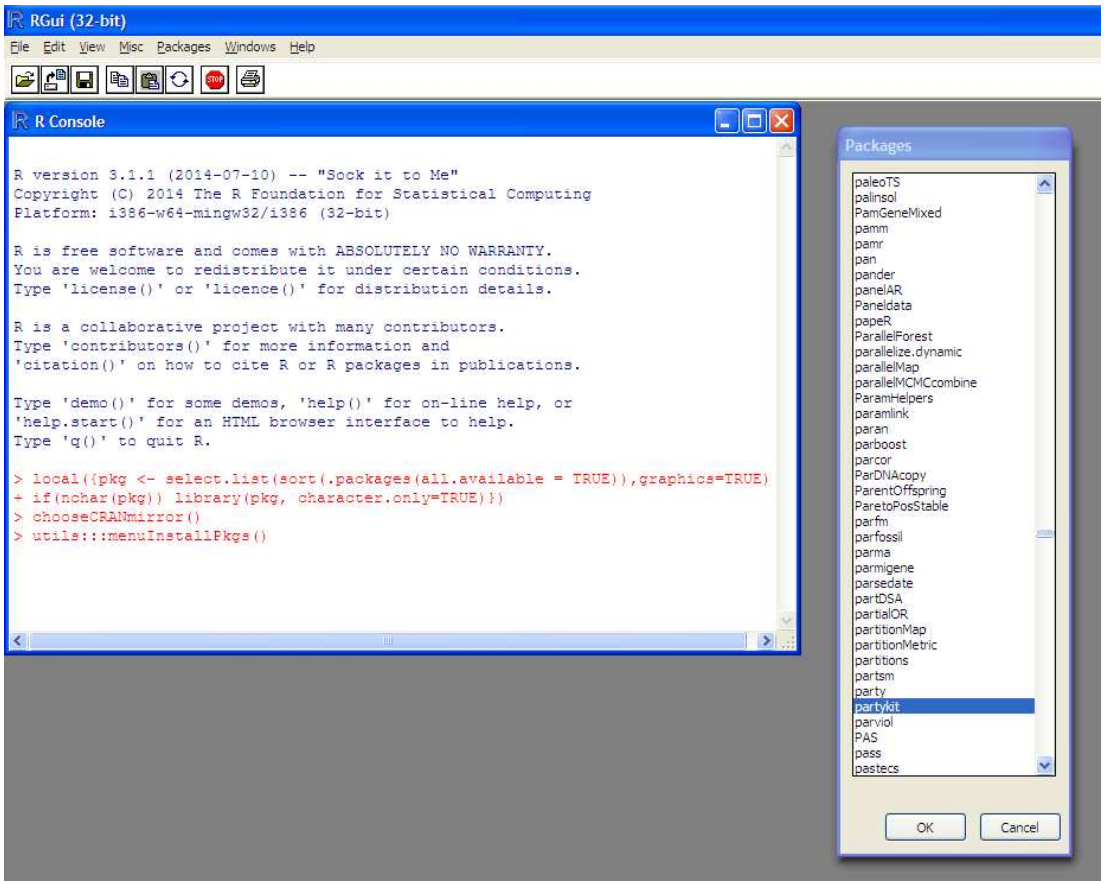


Ilustración 25: Carga de un paquete en R

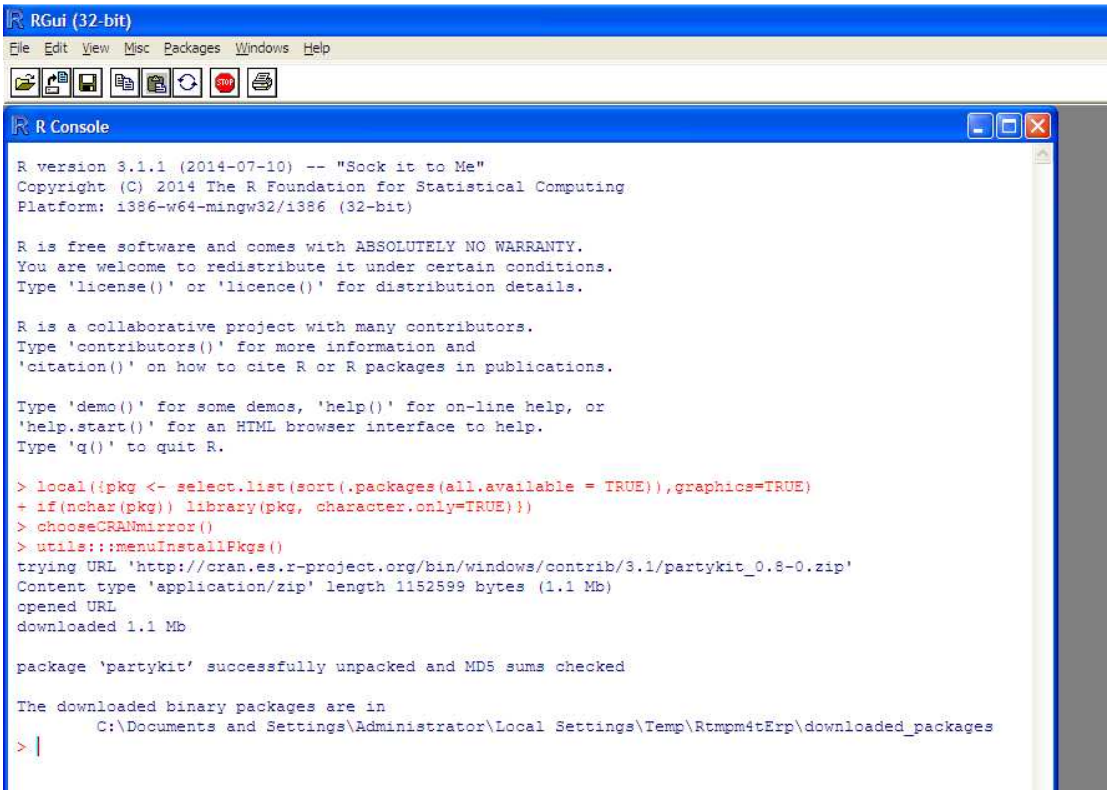


Ilustración 26: Paquete rpart

Package 'rpart'

June 27, 2012

Priority recommended

Version 3.1-54

Date 2012-06-27

DateNote March 2002 version of rpart

Author Terry M Therneau and Beth Atkinson <atkinson@mayo.edu>. R port by Brian Ripley. Note that maintainers are not available to give advice on using a package they did not author.

Maintainer Brian Ripley <ripley@stats.ox.ac.uk>

Description Recursive partitioning and regression trees

Title Recursive Partitioning

Depends R (>= 2.14.0), graphics, stats, grDevices

Ilustración 27: Intalación y carga del paquete rpart

Instalando y usando el paquete “rpart”:

- `install.packages('rpart',dependencies=TRUE)`
- `library(rpart)`

5.4. Modelo de regresión logística

En estadística, la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística.

En el análisis de regresión múltiple, la construcción, evaluación y selección del mejor subconjunto de variables predictoras que expliquen una variable respuesta es un problema importante de la estadística por diversas razones que incluyen:

I. Estimar o predecir a un menor costo al reducir el número de variables sobre las que se recogen datos.
II. Predecir con precisión mediante la eliminación de las variables sin relevancia.
III. Describir un conjunto de datos multivariados con parsimonia. Se dice que un modelo es parsimonioso si consigue ajustar bien los datos pero usando la menor cantidad de variables predictoras posibles.
IV. Estimar los coeficientes de regresión con errores estándar pequeños (sobre todo cuando algunas variables predictoras están altamente correlacionadas).
V. Emplear un menor conjunto de variables predictoras de forma que se mitigue el esfuerzo computacional.

Por lo anterior, el estudio de la selección del mejor subconjunto de variables no es un trabajo fácil, especialmente cuando se tiene un gran número de variables predictoras y no se tiene información precisa sobre la relación exacta entre las variables. A veces el número del total de posibles modelos es enorme, (2k, millones), es decir cuando existen más de k=20 variables predictoras, la evaluación de todas las posibles combinaciones de subconjuntos de variables es una tarea que puede tener un alto costo computacional. Por lo tanto, las técnicas de optimización combinatorial y las

estrategias para la selección de modelos tienen gran importancia y son necesarias para explorar el gran espacio de soluciones.

Las estrategias más conocidas y utilizadas para la selección del mejor subconjunto de variables son los métodos “Stepwise”, donde el procedimiento se basa en seleccionar el mejor modelo de manera secuencial incluyendo o excluyendo una sola variable predictora en cada paso según criterios de evaluación. Existen tres algoritmos usualmente usados: “Backward Elimination” (Eliminación hacia atrás), “Forward Selection” (Selección hacia adelante) y “Stepwise Selección” (Selección Paso a Paso).

Algunos criterios estadísticos basados en información del modelo como el criterio de información de Akaike (AIC), el criterio de información Bayesiano (BIC) y/o el criterio de información de Schwartz (SIC), que evalúan el grado de calidad de la regresión múltiple según el subconjunto de variables, presentan debilidad para medir la complejidad del modelo a partir del número de variables predictoras en términos de penalidad, la cual es una medida de compensación por el sesgo en la falta de ajuste cuando los estimadores de máxima verosimilitud son utilizados. Sin embargo, no es suficiente medir la complejidad del modelo (término penalidad, por ejemplo 2k en AIC) únicamente con variables predictoras o parámetros del mismo modelo. Es necesario considerar más elementos de juicio para definir y medir la complejidad de la información del modelo seleccionado. En este trabajo se propone utilizar una técnica de optimización combinatorial que sea computacionalmente eficiente en la selección del modelo estadístico con un criterio de evaluación que contenga más propiedades que relacionen e interactúen las componentes de un modelo de regresión.

5.4.1. Análisis de regresión: diagnósticos

Los diagnósticos de regresión se refieren a la clase general de técnicas para la detección de problemas en regresión, en el modelo o en los datos. Estos métodos están diseñados para detectar fallas en los supuestos, observaciones atípicas, deficiencias en el modelo y detección de situaciones en las que las relaciones fuertes entre las variables independientes están afectando los resultados. En este campo de

investigación se han hecho algunas publicaciones, sin embargo no existe una frontera clara entre la utilidad de estas técnicas con el tiempo. A continuación se mostrarán las técnicas para diagnósticos cuando se presentan problemas de multicolinealidad y para la detección de puntos influenciables.

5.4.1.1. Diagnósticos de colinealidad

El problema de colinealidad en regresión se refiere a que las columnas de la matriz de regresión X pueden estar casi linealmente dependientes o colineales, lo cual conlleva a que X'X esté cerca de ser singular. Entonces la matriz de varianza-covarianza (1) está cerca de la colinealidad, teniéndose un efecto considerable en la precisión. Luego si los coeficientes del modelo de regresión lineal (β) pueden ser estimados y tienen grandes varianzas, las pruebas de estimación de los parámetros del modelo tienen poca influencia y los intervalos de confianza podrían ser muy amplios, haciendo difícil decidir si una variable hace una contribución significativa a la regresión.

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1} \qquad (1)$$

A través del coeficiente de determinación múltiple (R^2) se puede detectar una relación dependiente cuando es cercano a 1 ó 100% para cada par de variables predictoras, sin embargo cuando existen outliers esta medida no es completamente apropiada. Por otro lado, cuando se extiende el caso a más de dos variables predictoras, un conjunto $(\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots\ldots \bar{X}_k)$ son colineales si para las constantes $(c_0, c_1, c_2, \ldots\ldots c_k,)$, la siguiente relación: $(c_1 * X_1 + c_2 * X_2 + \cdots \ldots\ldots c_k * X_k = c_0\ )$, se cumple y el R_k^2 de una variable de regresión X_k con las demás variables predictoras es cercano a 1, entonces se puede considerar que existe multicolinealidad.

Otra medida para detectar colinealidad es el factor de inflación de la varianza para el k-ésimo coeficiente de regresión (VIF_k). Consideramos el modelo de regresión lineal múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \qquad (2)$$

Entonces la varianza del k-ésimo coeficiente de regresión estimado es:

$$Var(\hat{\beta}) = \sigma^2 \left(\frac{1}{1-R_k^2} \right) \left(\frac{1}{S_{X_k X_k}} \right) \quad .. \quad (3)$$

La medida 1/(1- R_k^2) es denominada el k-ésimo factor de inflación de la varianza o (VIF_k) Si el valor de R_k^2 es cercano a 1 entonces la varianza de los parámetros estimados del modelo ($\hat{\beta}_k$) aumenta demasiado. En otras palabras, el VIF representa el incremento en la varianza del coeficiente de regresión estimado de una variable predictora debido a la presencia de colinealidad. Una variable predictora con un VIF mayor a 10, puede causar colinealidad. Para calcular los VIF se utiliza la inversa de la matriz de correlaciones C^{-1} y luego los VIF's serán los elementos de la diagonal principal C^{-1}.

5.4.1.2. Influencia estadística

En el análisis de regresión es importante realizar el diagnóstico de influencia estadística para analizar y conocer qué observaciones muestrales afectan en mayor grado el ajuste del modelo de regresión. En la literatura especializada, gran cantidad de autores se han enfocado en medidas de influencia proporcionando metodologías para evaluar el efecto en el ajuste del modelo y/o en el cambio de los coeficientes de regresión estimados al eliminar la i-ésima observación del conjunto de datos. Algunas de las más comunes medidas de influencia son: La distancia de Cook D_i, el $DFFITS_i$, el $DFBETA_{j(i)}$, el $COVRATIO_i$, la estadística Q_i, entre otros. La estadística Q_i permite evaluar para la i-ésima observación, el cambio en SCE cuando el modelo Y=Xβ+ε se ajusta después de eliminar dicha observación, es decir:

$$Q_i = \frac{\varepsilon_i^2}{(1-h_{ii})} = SCE - SCE(i) \quad \text{.....(4)}$$

Donde SCE es la suma de cuadrados residual cuando el modelo se ajusta con todas las n observaciones y $SCE_{(i)}$ es la suma de cuadrados residual cuando el modelo se ajusta sin la i-ésima observación.

5.4.2. Criterios estadísticos de selección de subconjuntos de variables

Los criterios estadísticos están basados en el principio de parsimonia, donde se recomienda seleccionar un modelo con la suma de los cuadrados residuales pequeños utilizando el mínimo número de variables. No obstante la selección del criterio puede dar lugar a diferentes opciones de tamaño del subconjunto de variables y pueden darse diversos puntos de vista de la magnitud de las diferencias entre los subconjuntos de modelos, siendo esto un aspecto relevante cuando se comparan modelos competentes Estos criterios pueden ser divididos en tres clases: Criterios de predicción, criterios de información o verosimilitud y criterios de maximización bayesiana con distribución a posteriori de probabilidad.

5.4.2.1. Criterio de información de Akaike: AIC

Akaike es uno de los pioneros en el campo de la evaluación de modelos estadísticos y aporta a la temática de selección de modelos el criterio de información de Akaike (AIC) definido como:

$$AIC = -2\log L(\hat{\theta}) + 2p \quad \text{.....(5)}$$

Donde L(θˆ) es la función de máxima verosimilitud y p es el número de parámetros en el modelo. El criterio precisa que el modelo con el menor valor AIC es seleccionado como el mejor al que se ajustan los datos. La estructura del AIC está compuesta entre la maximización del logaritmo de verosimilitud, es decir ($-2\log L(\hat{\theta})$), como componente de la falta de ajuste del modelo y p como el número de parámetros estimados dentro del modelo como componente de penalidad. La penalidad es una medida de la complejidad o compensación por el sesgo debido a la falta de ajuste cuando los estimadores de máxima verosimilitud son empleados. Para un modelo de regresión lineal múltiple ($Y = X \cdot \beta + \varepsilon$), el criterio de información de Akaike (AIC) se define a continuación:

$$AIC = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2(k+1) \quad \text{.....(6)}$$

Donde $(\hat{\sigma}^2)$ es la varianza de los residuales, k es el número de variables predictoras en el modelo de regresión y n es el número de observaciones de la muestra.

5.4.2.2. Criterio de información de Akaike corregido: AICc

En el criterio AIC definido en la ecuación (1), el sesgo es aproximado por el número de parámetros los cuales son constantes y no tienen variabilidad. Para el modelo de regresión múltiple, la corrección del sesgo del logaritmo de la verosimilitud es calculada como:

$$Sesgo = E_G\left[\frac{1}{n}\sum_{i=1}^{n}\log f(x_i \mid \hat{\theta}) - \int_R \log f(X \mid \hat{\theta}) dG(X)\right] = \frac{n(k+1)}{n-k-2} \quad \text{..... (7)}$$

Si se emplea la ecuación (6) para el modelo de regresión múltiple, se puede definir el criterio AICc para una muestra finita, el cual fue propuesto originalmente por, como se observa a continuación:

$$AIC_C = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2\frac{n(k+1)}{n-k-2} \quad \text{.....(8)}$$

De manera similar que en el AIC, se selecciona el modelo con el menor valor AICc.

5.4.2.3. Criterio de información bayesiano: BIC

Para mejorar la inconsistencia del criterio AIC, Akaike y Schwarz presentaron un criterio de selección de modelos desde la perspectiva bayesiana. Schwarz estableció que la solución de bayes consiste en seleccionar el modelo con una alta probabilidad a posteriori. Para grandes muestras esta probabilidad a posteriori puede ser aproximada por la expansión de Taylor. Schwarz define el primer término de su criterio como el logaritmo de los estimadores de máxima verosimilitud (MLE's) para el modelo y el

segundo término como p*log(n), entonces el criterio de información bayesiano (BIC) es definido como sigue:

$$BIC = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + p\log(n) \quad \text{.....(9)}$$

Donde p es el número de parámetros en el modelo y n es el tamaño de muestra. El criterio selecciona el mejor modelo como el que tiene el menor valor BIC.

5.4.2.4. Criterio de complejidad de la información: ICOMP

Bozdogan presenta un nuevo criterio de selección del mejor modelo estadístico basado en la definición de complejidad de un modelo, describiéndola en términos de la interacción entre componentes de un modelo y la información pertinente para su construcción. Luego presenta el enfoque de complejidad de la información ICOMP (IFIM) para la evaluación de modelos basado en la complejidad de máxima covarianza. Finalmente para un modelo lineal normal multivariado o no lineal, presenta el criterio de selección de modelos ICOMP (IFIM) definido en forma general como:

$$ICOMP\,(IFIM) = -2\log L(\hat{\theta}) + 2C_1(F^{-1}(\hat{\theta})) \quad \text{.....(10)}$$

Donde, C_1 define la máxima complejidad de la información de F^{-1} expresada como la matriz inversa de información estimada de Fisher (IFIM) de un modelo, o conocida como la matriz de límite inferior de Crámer-Rao (CRLB). El enfoque de ICOMP (IFIM) aprovecha las propiedades asintóticas óptimas de los estimadores de máxima verosimilitud y utiliza la información basada en la complejidad de la matriz inversa de información de Fisher (IFIM). Finalmente, el criterio ICOMP (IFIM) para un modelo de regresión múltiple, selecciona el mejor modelo como el que tiene el menor valor ICOMP (IFIM) definido de la siguiente forma:

$$ICOMP_{IFIM} = nlog(2\pi) + nlog\left(\widehat{\sigma^2}\right) + n + \cdots$$

$$\ldots\ (q+1)log\left(\frac{Tr\left(\sigma^2(X'X)^{-1}\right)+\frac{2\widehat{\sigma^4}}{n}}{q+1}\right) - \ldots$$

$$\ldots\ log\left|\widehat{\sigma^2}(X'X)^{-1}\right| - log\left(\frac{2\widehat{\sigma^4}}{n}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (11)$$

5.4.3. Métodos estadísticos de selección de subconjuntos de variables

5.4.3.1. Método de selección backward elimination: SBS

El método inicia con el modelo completo (todas las k variables predictoras). En cada paso se va eliminando una variable del modelo según se cumpla una de las siguientes condiciones:

1) La variable con el menor valor del estadístico F parcial definido como:

$$F_p = \frac{SSR_k - SSR_{k-1}}{MSE_k} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

Donde SSR_k es la suma de cuadrado de la regresión con k variables, SSR_{k-1} es la suma de cuadrados de la regresión con k-1 variables y MSE_k es el cuadrado medio del error del modelo con las k variables. Se calcula el F_p para cada una de las variables que se encuentren en el modelo y se excluye la variable que tiene el F_p más pequeño.

2) La variable que genera la menor reducción en el R^2 al ser descartada del modelo.

3) La variable que tiene el menor coeficiente de correlación parcial en valor absoluto con la variable dependiente.

El proceso del método finaliza cuando se llega a un número prefijado p* de variables predictoras o cuando el valor del F_p de todas las variables no eliminadas en el modelo es mayor a un valor fijado F_{out} (F_{out} usualmente 4). Es común fijar con anterioridad un nivel de significancia dado α* (por lo general del 10%) para la prueba “t” o “F” en cada paso y termina el método cuando todos los valores p son menores que α*. El

inconveniente en este método es que una variable que ha sido eliminada del modelo, nunca puede entrar en la regresión de nuevo.

5.4.3.2. Método de selección forward selection : SFS

Este método inicia con un modelo que tiene solo el término constante (ε). Se utiliza la variable predictora con mayor correlación con la variable dependiente en valor absoluto. Si la primera variable no es significativa entonces se tiene el modelo $\hat{Y} = \bar{Y}$ y se detiene el proceso, sino la siguiente variable que entra al modelo cumple con cualquiera de las siguientes condiciones:

1) La variable que tiene el mayor F_p entre las variables que no están incluidas en el modelo.

2) La variable que genera el mayor crecimiento del R^2 al ser incluida en el modelo.

3) La variable que tiene el mayor coeficiente de correlación parcial en valor absoluto con la variable dependiente.

El proceso de este método finaliza cuando se obtiene un número fijado p* de variables predictoras o cuando el valor del F_p de todas las variables que aún no han sido incluidas en el modelo es menor a un valor fijado F_{in} (F_{in} usualmente igual a 4). Es común fijar con anterioridad un nivel de significancia dado α* (por lo general del 5%) para la prueba “t” o “F” en cada paso y termina el método cuando todos los valores p de las variables no incluidas son aún mayores que α*. El problema en este método es que una variable que ha sido incluida en el modelo, nunca puede ser removida de la regresión.

5.4.3.3. Método de selección stepwise selection: SS

Este método propuesto por Efroymson y Draper y Smith, combina los métodos SFS y SBS y también es conocido como algoritmo de regresión por pasos (stepwise regression algorithm), en el cual comienza con el SFS seguido por el SBS en cada

paso. Este algoritmo inicia con el modelo que contiene solo el término constante (ε) y enseguida ejecuta el paso SFS adicionando una sola variable. Luego se aplica el paso SBS el cual remueve una variable si el correspondiente F_p es menor que el F_{out}. Es de notar que en este algoritmo se usan F_{in} y F_{out} con $F_{in} \leq F_{out}$. Esta combinación se repite hasta que ninguna de las variables que no han sido seleccionadas, tengan el grado de importancia necesaria como para ser incluidas en el modelo.

5.5. Modelos de máquina de vector soporte

Las máquinas de soporte vectorial o máquinas de vectores de soporte (Support Vector Machines, SVM_s) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T.

Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase.

Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta.

Dado un conjunto de puntos, subconjunto de un conjunto mayor (espacio), en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo (cuya categoría desconocemos) pertenece a una categoría o a la otra.

Como en la mayoría de los métodos de clasificación supervisada, los datos de entrada (los puntos) son vistos como un vector p-dimensional (una lista de p números).

La SVM busca un hiperplano que separe de forma óptima a los puntos de una clase de la de otra, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior.

En ese concepto de "separación óptima" es donde reside la característica fundamental de las SVM: este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. Por eso también a veces se les conoce a las SVM como clasificadores de margen máximo. De esta forma,

los puntos del vector que son etiquetados con una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado.

Los algoritmos SVM pertenecen a la familia de los clasificadores lineales. También pueden ser considerados un caso especial de la regularización de Tikhonov.

En la literatura de los SVMs, se llama atributo a la variable predictora y característica a un atributo transformado que es usado para definir el hiperplano. La elección de la representación más adecuada del universo estudiado, se realiza mediante un proceso denominado selección de características.

Al vector formado por los puntos más cercanos al hiperplano se le llama vector de soporte.

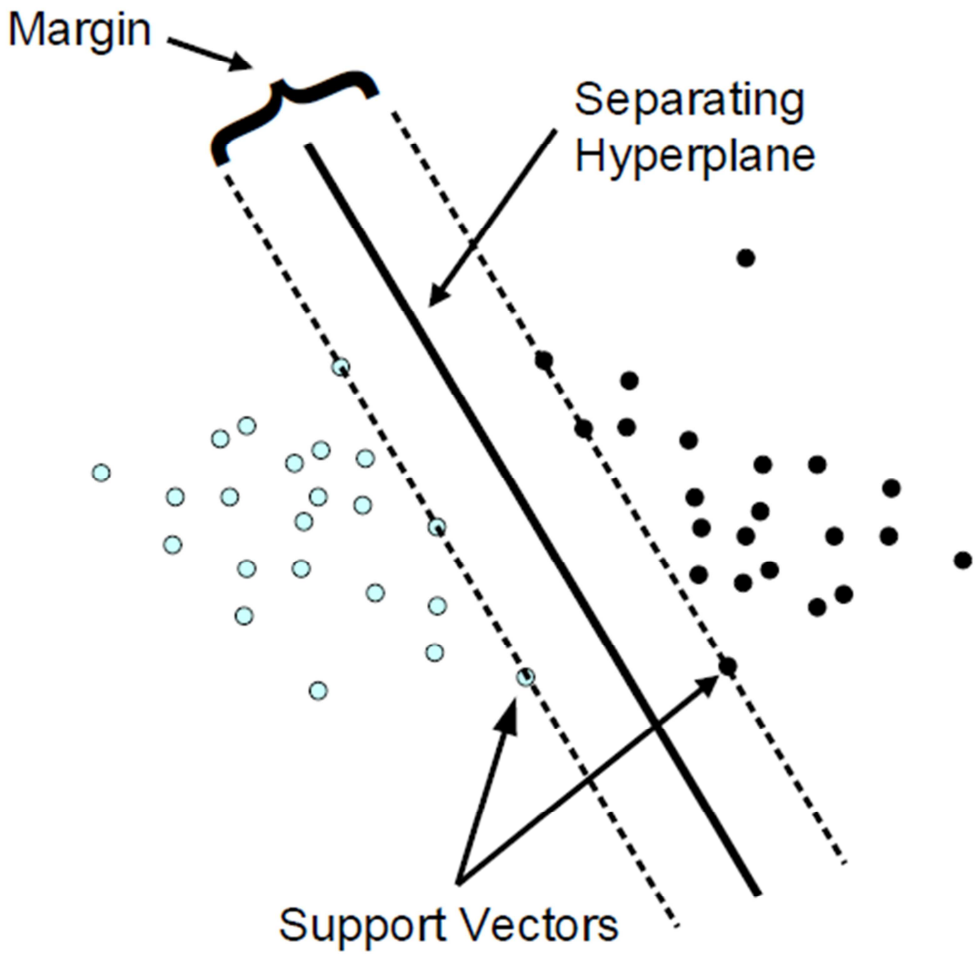
Los modelos basados en SVMs están estrechamente relacionados con las redes neuronales. Usando una función kernel, resultan un método de entrenamiento alternativo para clasificadores polinomiales, funciones de base radial y perceptrón multicapa.

Dado un conjunto de datos de entrenamiento$\{x_i, y_i\}_{i=1}^{n} \in R^m \times \{\pm 1\}$, se desea encontrar el hiperplano óptimo que divida las dos clases de datos. El correspondiente hiperplano puede ser definido como:

$$W^T x + b = 0$$

donde x es un vector de datos, w^Tes el vector de parámetros de nuestro modelo y b es un término independiente que ofrece mayor libertad al momento de encontrar el hiperplano óptimo para clasificar datos.

Ilustración 28:Hiperplano para un caso linealmente separable



Dado un conjunto de puntos linealmente separables, ilustrados en la figura x como cruces y círculos, se puede usar la distancia r para calcular un margen de separación ρ, de la siguiente manera:

$$\rho = 2r * = \frac{2}{\|w\|}$$

Así para asegurar encontrar el hiperplano, se minimiza ρ con respecto a x y b:

$$min \frac{1}{2} \|w\|^2$$

con la restricción $y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$

Los puntos que aparecen en la figurax que se encuentra sobre las líneas no punteadas se les conoce como vectores de apoyo.

Cuando los datos de prueba no son linealmente separables se puede adoptar dos técnicas para resolverlo: con optimización " margen suave" y a través de kernel. Para el primer método mencionado se agrega una variable τ_i, la cual es usada para registrar la cantidad de violación de clasificación de un clasificador, quedando la ecuación como sigue:

$$min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \tau_i$$

con la restricción $y_i(w^T x_i + b) \geq 1 - \tau_i, \tau_i \geq 0,\ i = 1, \ldots, n,$

al utilizar Kernel se debe obtener el clasificador óptimo

$$f(x) = \sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b$$

donde α es el multiplicador óptimo de Lagrange y $K(x_i, x)$ es una función kernel. Los Kernels comúnmente usados son:

1. Polinomial $K(x_i, x) = (x_i^T \cdot x + c)^d$
2. Función de base radial $K(x_i, x) = exp(-\gamma \|x_i - x\|^2), \gamma > 0$
3. Sigmoidal $K(x_i, x) = tanh(x_i^T \cdot x + c)$

5.6. Modelo de particionado recursivo

El particionamiento recursivo es una técnica estadística de análisis multivariante. Su objetivo es el de construir árboles de decisión que modelen la influencia de una serie de variables explicativas sobre la variable objetivo de un estudio estadístico.

En función de la naturaleza, discreta o continua, de la variable objetivo, los árboles construidos suelen denominarse árboles de clasificación o de regresión.

Los modelos construidos con dicha técnica rivalizan con otros más tradicionales de la estadística —por ejemplo, regresiones logísticas— o de la inteligencia artificial, como los basados en redes neuronales.

Esta técnica fue introducida por Leo Breiman en 1984.1 Hoy en día existen diversas implementaciones de estas técnicas que desarrollan el concepto original y diversos paquetes estadísticos son capaces de construir árboles basados en dichos principios.

La implementación original del algoritmo de los autores es mantenida y desarrollada por Salford Systems y tiene el nombre comercial de CART.

El lenguaje de programación R incluye el paquete rpart que permite construir dichos modelos.

En comparación con otros métodos multivariables, el particionamiento recursivo tiene ventajas y desventajas.

Las ventajas son:

- Genera modelos más intuitivos que no requiere que el usuario realice cálculos.
- Permite variar priorización de errores en la clasificación con el fin de crear una regla de decisión que tiene más sensibilidad o especificidad.
- Puede ser más preciso.

Las desventajas son:

- No funciona bien para las variables continuas

- Puede haber overfit datos.

En nuestro caso, nos vamos a centrar en dar una breve muestra de los métodos que se encuentran en las rutinas rpart y que posteriormente han sido implementados en R para desarrollar el modelo propuesto.

Ilustración 29: Rutina Rpart 1

1. Grow the Tree

To grow a tree, use

rpart(*formula***, data=, method=,control=)** where

formula	is in the format *outcome ~ predictor1+predictor2+predictor3*+ect.
data=	specifies the data frame
method=	**"class"** for a classification tree **"anova"** for a regression tree
control=	optional parameters for controlling tree growth. For example, control=rpart.control(minsplit=30, cp=0.001) requires that the minimum number of observations in a node be 30 before attempting a split and that a split must decrease the overall lack of fit by a factor of 0.001 (cost complexity factor) before being attempted.

Ilustración 30: Rutina Rpart 2

2. Examine the results

The following functions help us to examine the results.

printcp(*fit*)	display cp table
plotcp(*fit*)	plot cross-validation results
rsq.rpart(*fit*)	plot approximate R-squared and relative error for different splits (2 plots). labels are only appropriate for the "anova" method.
print(*fit*)	print results
summary(*fit*)	detailed results including surrogate splits
plot(*fit*)	plot decision tree
text(*fit*)	label the decision tree plot
post(*fit*, file=)	create postscript plot of decision tree

Ilustración 31: Rutina Rpart 3

3. prune tree

Prune back the tree to avoid overfitting the data. Typically, you will want to select a tree size that minimizes the cross-validated error, the **xerror** column printed by **printcp( )**.

Prune the tree to the desired size using
prune(*fit*, cp=)

Specifically, use **printcp( )** to examine the cross-validated error results, select the complexity parameter associated with minimum error, and place it into the **prune( )** function. Alternatively, you can use the code fragment

```
fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]
```

to automatically select the complexity parameter associated with the smallest cross-validated error. Thanks to HSAUR for this idea.

Ilustración 32: Ejemplo árbol de decisión para clasificación 1

Classification Tree example

Let's use the data frame **kyphosis** to predict a type of deformation (kyphosis) after surgery, from age in months (Age), number of vertebrae involved (Number), and the highest vertebrae operated on (Start).

```
# Classification Tree with rpart
library(rpart)

# grow tree
fit <- rpart(Kyphosis ~ Age + Number + Start,
   method="class", data=kyphosis)

printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# plot tree
plot(fit, uniform=TRUE,
   main="Classification Tree for Kyphosis")
text(fit, use.n=TRUE, all=TRUE, cex=.8)

# create attractive postscript plot of tree
post(fit, file = "c:/tree.ps",
   title = "Classification Tree for Kyphosis")
```

Ilustración 33:Ejemplo árbol de decisión para clasificación 2

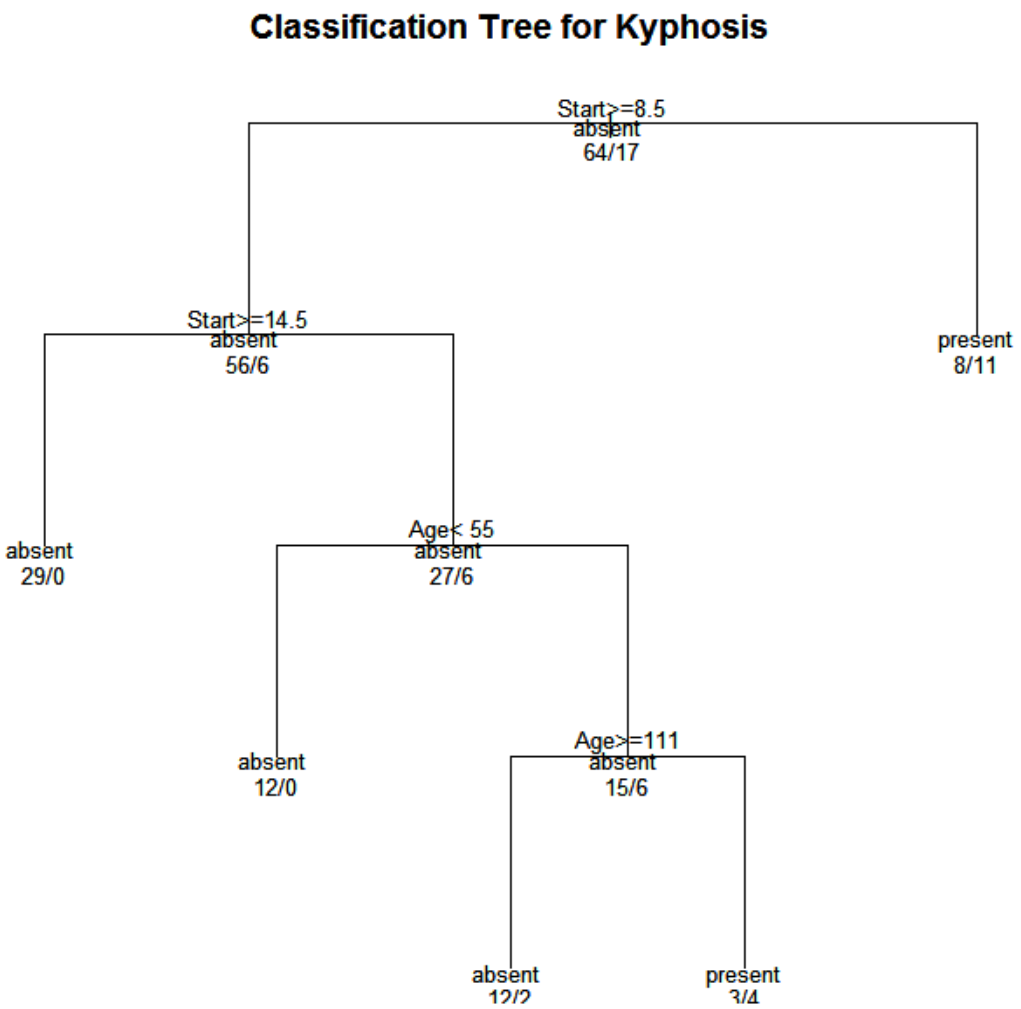


Ilustración 34: Ejemplo árbol de decisión para clasificación 3

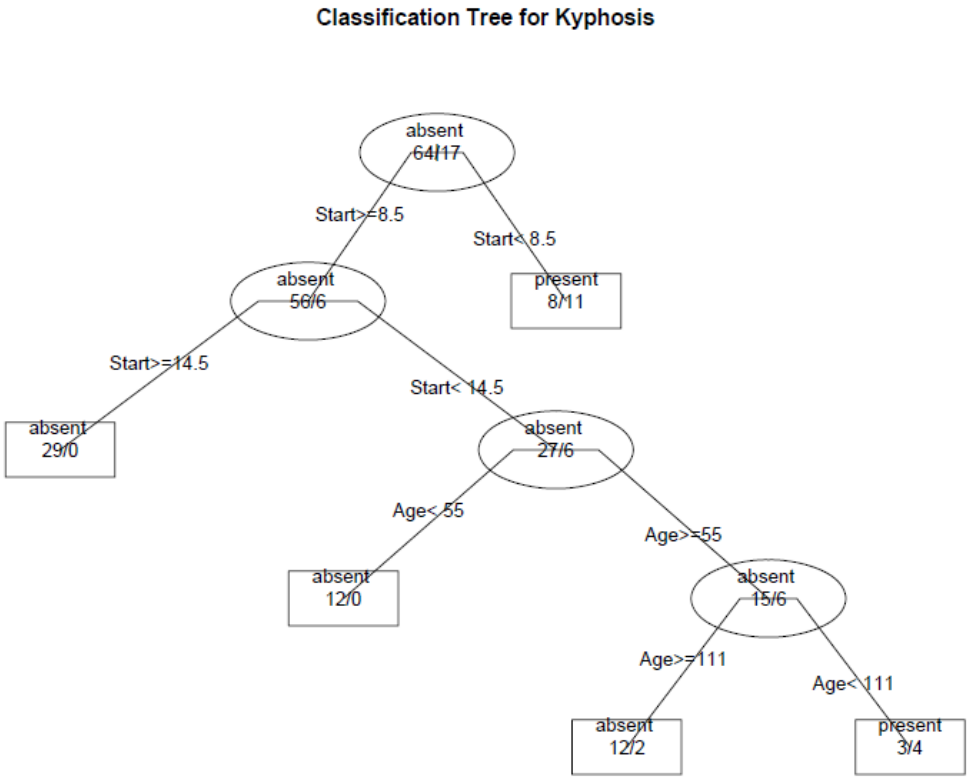


Ilustración 35: Ejemplo árbol de decisión para clasificación 4: Poda

```
# prune the tree
pfit<- prune(fit, cp=
  fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])

# plot the pruned tree
plot(pfit, uniform=TRUE,
   main="Pruned Classification Tree for Kyphosis")
text(pfit, use.n=TRUE, all=TRUE, cex=.8)
post(pfit, file = "c:/ptree.ps",
   title = "Pruned Classification Tree for Kyphosis")
```

Ilustración 36: Ejemplo árbol de decisión para clasificación 5: Árbol podado

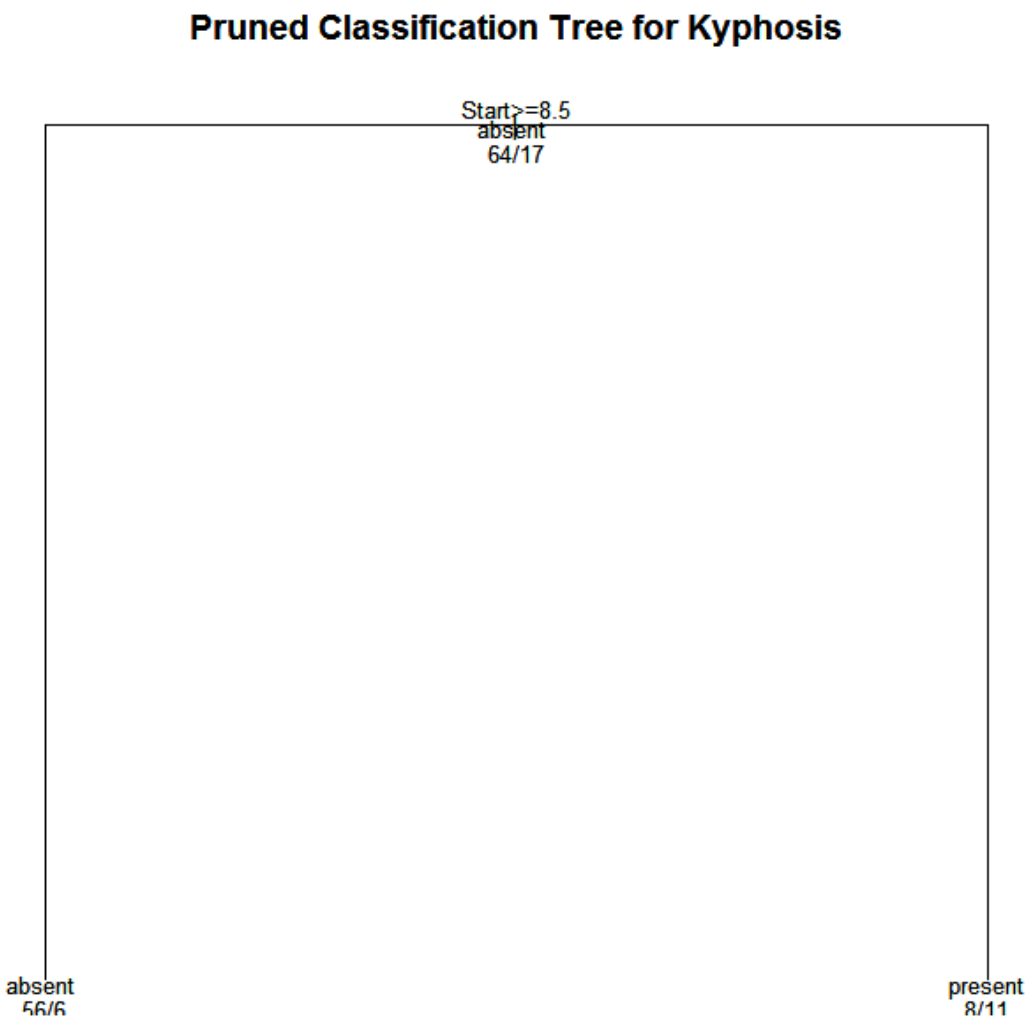


Ilustración 37: Ejemplo árbol de regresión 1

Regression Tree example

In this example we will predict car mileage from price, country, reliability, and car type. The data frame is **cu.summary**.

```
# Regression Tree Example
library(rpart)

# grow tree
fit <- rpart(Mileage~Price + Country + Reliability + Type,
   method="anova", data=cu.summary)

printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# create additional plots
par(mfrow=c(1,2)) # two plots on one page
rsq.rpart(fit) # visualize cross-validation results

# plot tree
plot(fit, uniform=TRUE,
   main="Regression Tree for Mileage ")
text(fit, use.n=TRUE, all=TRUE, cex=.8)

# create attractive postcript plot of tree
post(fit, file = "c:/tree2.ps",
   title = "Regression Tree for Mileage ")
```

Ilustración 38: Ejemplo árbol de regresión 2

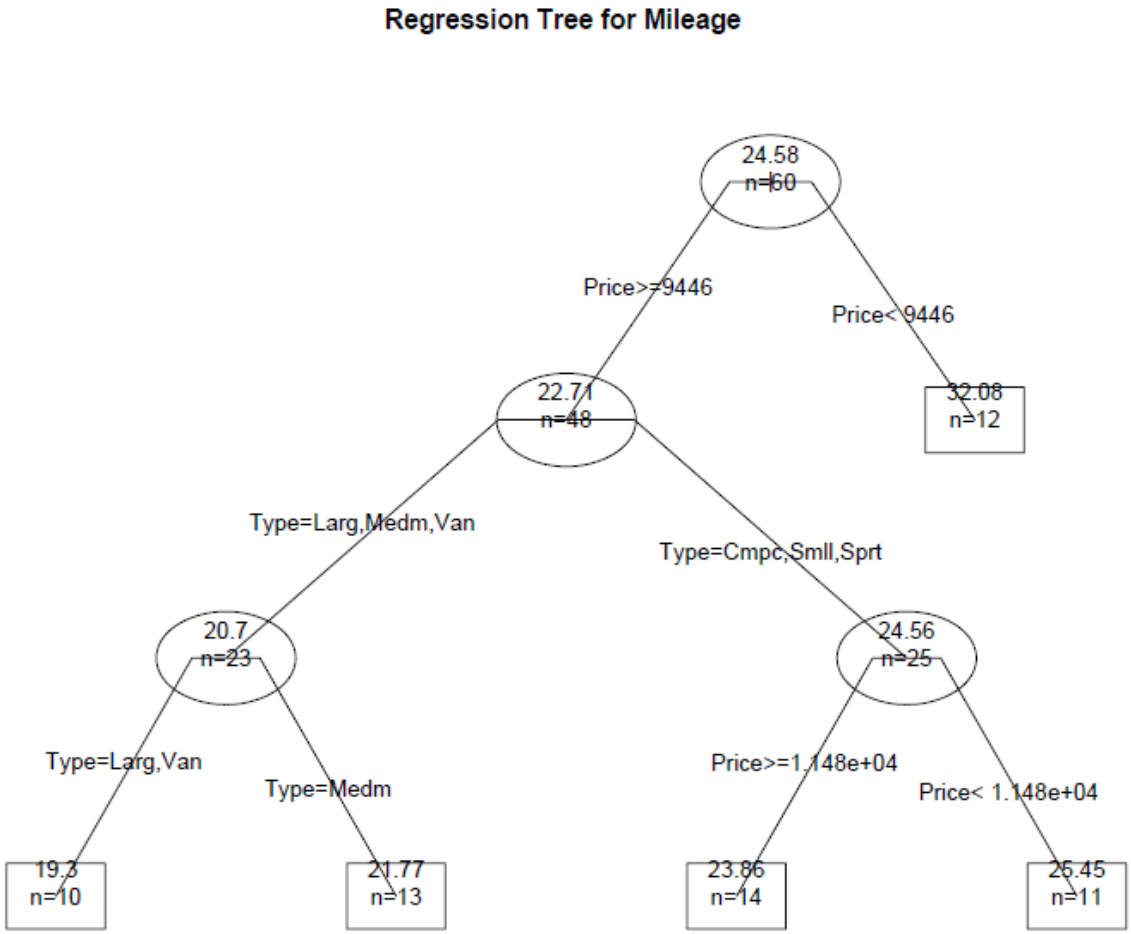


Ilustración 39:Ejemplo árbol de regresión 3

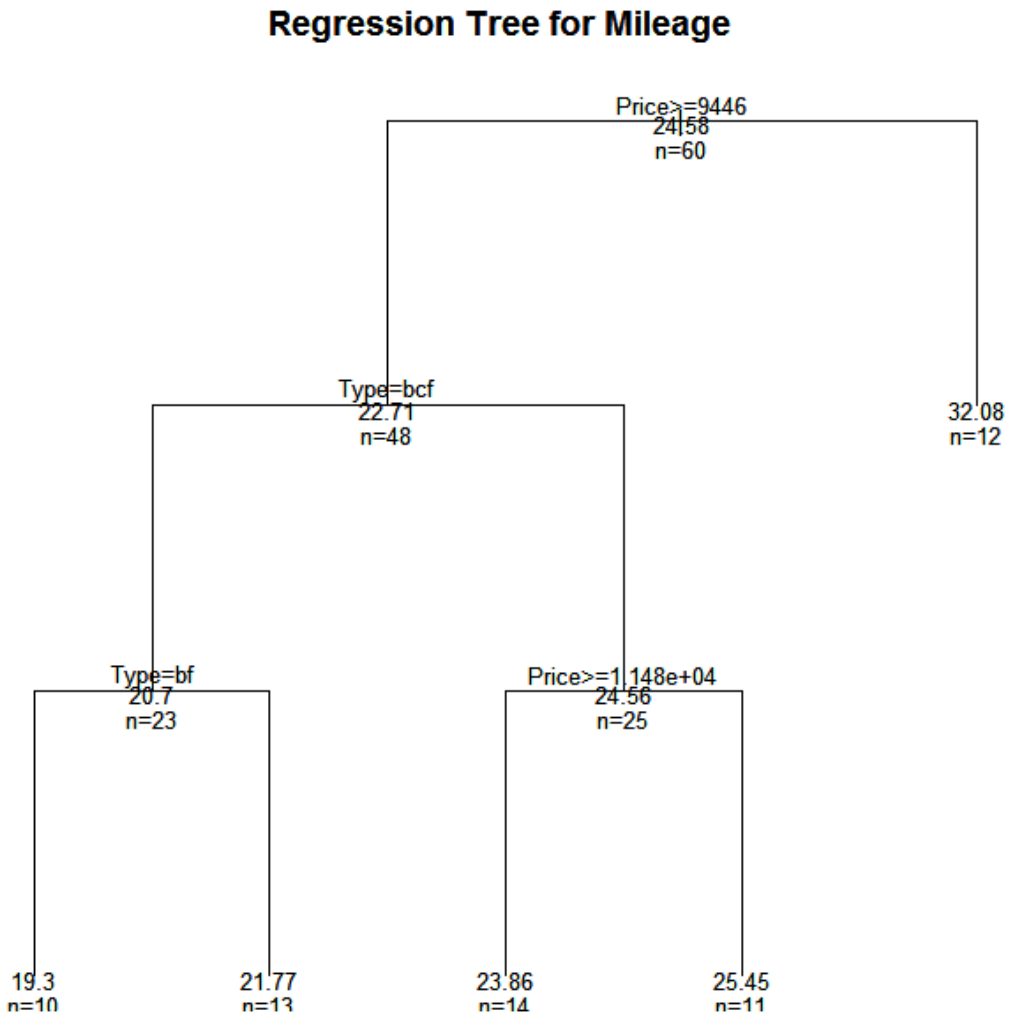


Ilustración 40: Ejemplo árbol de regresión 4: Poda

```
# prune the tree
pfit<- prune(fit, cp=0.01160389) # from cptable

# plot the pruned tree
plot(pfit, uniform=TRUE,
   main="Pruned Regression Tree for Mileage")
text(pfit, use.n=TRUE, all=TRUE, cex=.8)
post(pfit, file = "c:/ptree2.ps",
   title = "Pruned Regression Tree for Mileage")
```

.

6. Pruebas y resultados

Para el desarrollo del presente trabajo, se procede a recopilar una muestra de estadísticos descriptivos compuesta por la documentación contable y financiera de 16.398 empresas.

La fuente de información de donde procede la documentación es la base de datos SABI, ya introducida en anteriores apartados.

Se ha estimado la insolvencia de 2012 con información económico-financiera de 2011, y la insolvencia de 2011 con información económico financiera de 2010.

No se ha estimado el modelo para insolvencias de 2013 por carecer de datos actualizados para la totalidad de las empresas.

Las variables utilizadas para el estudio son las siguientes:

- Estado: Activa/Inactiva
- Tesorería
- Activo total
- Fondos Propios
- Importe neto de cifra de ventas
- Resultado de explotación
- Resultado del ejercicio
- Gastos de personal
- Gastos de amortización

y el ratio

- Fondos Propios/Activo Total

En la siguiente tabla se resumen los estadísticos descriptivos tenidos en cuenta para la elaboración de los modelos.

De partida se han considerado inactivas empresas con el estadístico Fondos Propios ≤ 0. Sólo con este descriptivo un modelo simple de estado acertaría la insolvencia empresarial con una probabilidad del 76%, luego cualquier modelo implementado debe superar el anterior resultado.

Tabla 26: Muestras estadísticos descriptivos

```
> head(datos)
  Estado Tes    TA    FP Ventas ResExp ResEje Gpers Gamort FonProp.TotAct
1      1  13    59     4    197    -61    -47   152      1              7
2      0  11 11138  5938    539   -538   -405   446    356             53
3      0  28  4072    10   2514    106     58   428     58              0
4      0  25  1363   268    886    -21    -55   168      3             20
5      0 118 16153 -1968  14115  -2957  -3149  7468   1036            -12
6      0  22  4076   486   3211    -23    -85   484     62             12
```

Miles de euros — Tanto por cien

```
> table(datos$Estado)

    0     1
 3921 12477
```

Se han considerado como inactivas las empresas con Fondos Propios <=0

Un modelo naïve acertaría con una probabilidad del 76%

Total empresas: 16.398

Activas FP>0: 12.477 Inactivas FP<0: 3.921

Una simple clasificación con un modelo univariante, acertaría en un 76% los casos de estado de solvencia / insolvencia de una muestra a clasificar

El objetivo de los modelos predictivos es la alerta temprana de situaciones de insolvencia empresarial (Concurso de acreedores, disolución, extinción, ….), base para cuantificar el riesgo crediticio y para ello tomamos como información los registros contables de la empresa.

La siguiente tabla nos muestra el flujo de información:

Ilustración 41: Flujo de información

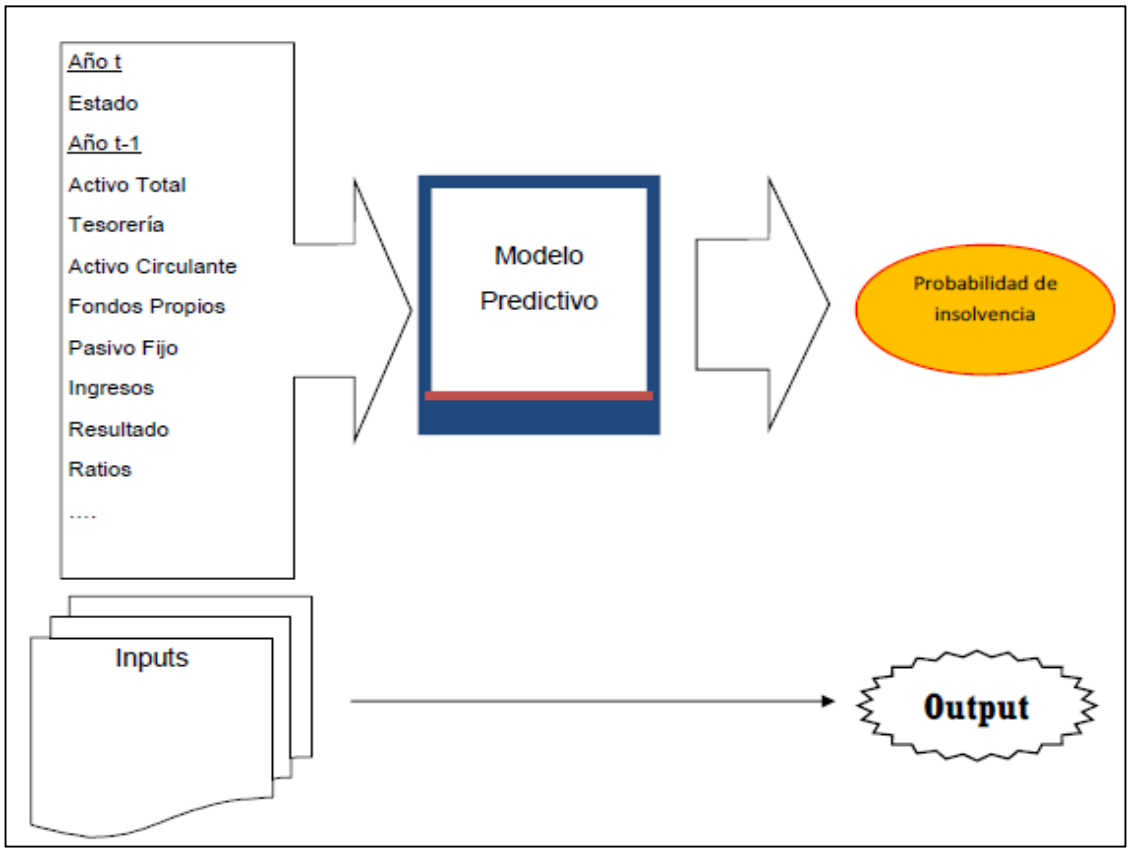
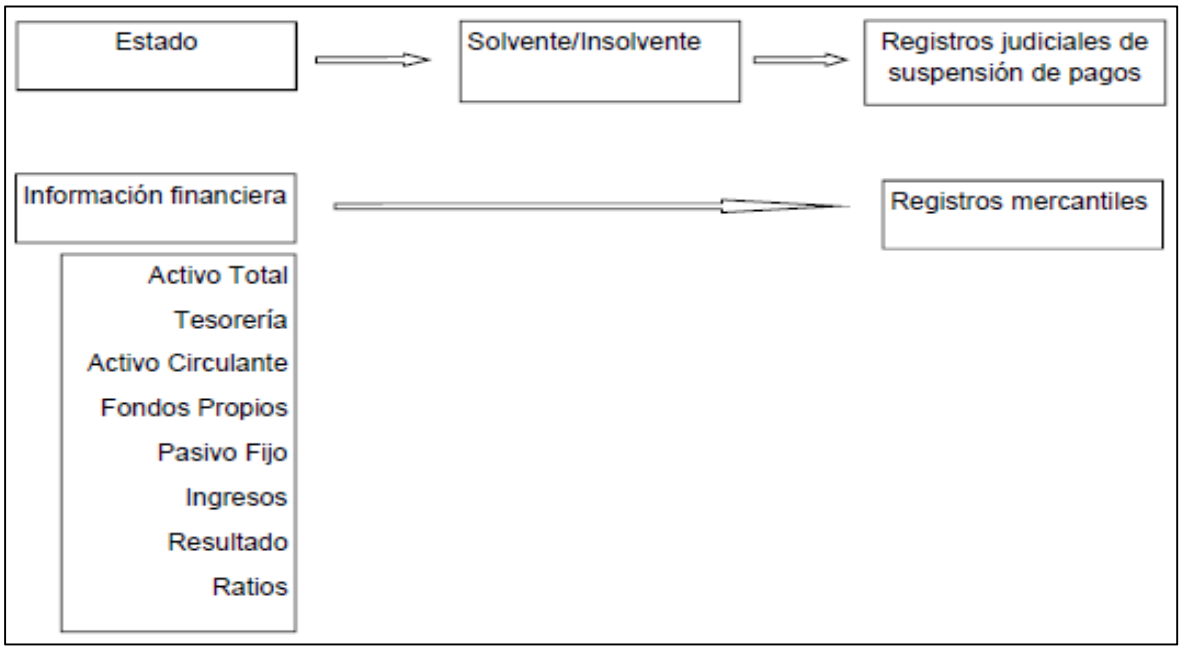


Ilustración 42: Fuentes de información



El modelo infiere el nivel más probable de insolvencia empresarial realizando clusters de empresas según el diferente grado de insolvencia estimada.

El número de clusters es parametrizable en función del tamaño muestral.

Tabla 27: Probabilidad de insolvencia

Cluster	Nº empresas	Probabilidad insolvencia
Máxima solvencia	3.171	4,2%
Solvencia alta	410	11,8%
Solvencia media-alta	213	19,8%
Solvencia media	102	34,3%
Solvencia media-baja	153	51,0%
Solvencia baja	86	79,1%
Insolvencia	1.365	100,0%
Total empresas	5.500	32,1%

La pertenencia a uno u otro clúster se establece a partir de una variable decisional derivada a partir de la información económico-financiera.

Los clusters pueden agruparse, según la semejanza de su probabilidad de insolvencia.

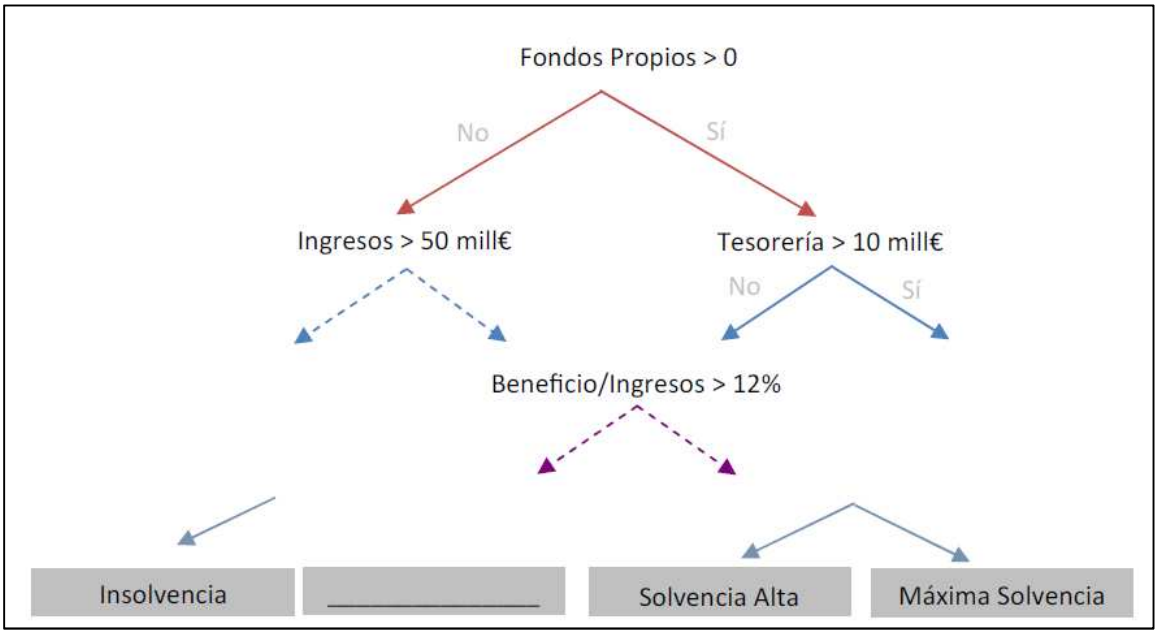
Por ejemplo el clúster de máxima Solvencia se formaría a partir del conjunto de las siguientes reglas:

- Fondos Propios positivos
- Tesorería > 10 mill €
- Resultado del ejercicio > 52,5 mill €
- Ratio Beneficio / Fondos Propios > 26,5%
- Ratio Beneficio / Cifra de negocios > 23,5%

Sólo el 4,2% de las 3.171 empresas que cumplían estas características resultaron ser insolventes un año después.

A partir de esto ya generamos el árbol decisional:

Ilustración 43: Generación del árbol decisional



La predicción de la insolvencia y la gestión de los clusters puede llevarse a cabo de forma segregada teniendo en cuenta, por ejemplo, el tamaño de la empresa o la actividad empresarial.

Resultado de muestra analizada para empresas con un activo no superior al millón de euros.

Tabla 28: Empresas con activo no superior al millón de euros

	datos$Estado.pred		
datos$Estado	0	1	Row Total
0	1546	222	1768
	1814.734	816.009	
	0.874	0.126	0.321
	0.906	0.059	
	0.281	0.040	
1	160	3572	3732
	859.713	386.576	
	0.043	0.957	0.679
	0.094	0.941	
	0.029	0.649	
Column Total	1706	3794	5500
	0.310	0.690	

Cluster	Nº empresas	Probabilidad insolvencia
Máxima solvencia	3.171	4,2%
Solvencia alta	410	11,8%
Solvencia media-alta	213	19,8%
Solvencia media	102	34,3%
Solvencia media-baja	153	51,0%
Solvencia baja	86	79,1%
Insolvencia	1.365	100,0%
Total empresas	5.500	32,1%

Resultados particulares de predicciones del modelo aplicado a empresas reales:

1. ROOM MATE, S.L. B82559261

Cuentas	2011	2010
Tesorería	0,48	0,23
Fondos Propios	-1,14	3,04
Total Activo	48,57	47,19
Pasivo Fijo	36,46	33,81
Ingresos Explotación	15,87	15,27
Resultado Ejercicio	-4,19	-4,34

Principales magnitudes Millones de euros

Probabilidad estimada de insolvencia: 100%

2. SIGLA, S.A. (GRUPO VIPS) A28308484

Cuentas	2012	2011
Tesorería	6,99	19,39
Fondos Propios	20,81	41,28
Total Activo	151,66	171,83
Pasivo Fijo	46,42	30,76
Ingresos Explotación	245,59	273,86
Resultado Ejercicio	-20,57	-29,19

Principales magnitudes Millones de euros

Probabilidad estimada de insolvencia: 78%

3. GRUPO VIAJES BARCELÓ B07012107

Cuentas	2012	2011
Tesorería	4,29	2,27
Fondos Propios	26,48	31,14
Total Activo	122,75	118,30
Pasivo Fijo	28,17	29,83
Ingresos Explotación	43,55	432,56
Resultado Ejercicio	-4,65	7,58

Principales magnitudes Millones de euros

Probabilidad estimada de insolvencia: 60%

4. FOOD SERVICE, PROJECT (GRUPO ZENA) B82798943

Cuentas	2012	2011
Tesorería	7,12	0,02
Fondos Propios	22,31	24,42
Total Activo	201,85	166,06
Pasivo Fijo	114,95	101,82
Ingresos Explotación	160,89	30,05
Resultado Ejercicio	6,30	1,39

Principales magnitudes Millones de euros

Probabilidad estimada de insolvencia: 42%

5. CADOR IBERIA, S.A. A28618676

Cuentas	2011	2010
Tesorería	0,09	0,10
Fondos Propios	3,92	3,99
Total Activo	9,27	8,59
Pasivo Fijo	1,56	4,29
Ingresos Explotación	1,16	1,50
Resultado Ejercicio	-0,06	0,00

Principales magnitudes Millones de euros

Probabilidad estimada de insolvencia: 6%

Se plantean tres modelos para la predicción de la insolvencia y posteriormente se compararán los resultados.

Para proceder al ajuste del modelo se van a realizar dos pasos:

1. *Training set*.
2. *Test set and simulation*.

Para ello y con objeto de sobreentrenar el modelo (elevada capacidad explicativa y reducida capacidad predictiva), se ha particionado la muestra en dos grupos.

Para *el training set* se utilizará el 80% de la muestra y para el *test set* el resto.

En el proceso de simulación, se han llevado a cabo 5.000 simulaciones para poder analizar la robustez de los resultados.

6.1. Modelo de regresión logística

Como ya se ha visto en el apartado de desarrollo, en primer lugar se obtiene una función que liga la variable dependiente (Estado) con el conjunto de variables independientes.

La elección de las variables se hace mediante un análisis stepforward según el criterio Akaike information criterium (AIC).

Si bien la variable Estado es dicotómica, el resultado de la función logística no tiene porqué serlo.

En la siguiente tabla vemos los resultados arrojados por el modelo y el estado predecido.

Tabla 29: Resultado modelo de Regresión Logística

```
> head(test)
   Estado Tes    TA   FP Ventas ResExp ResEje Gpers Gamort FonProp.TotAct new.Estado Estado.pred
4       0  25  1363  268    886    -21    -55   168      3             20  0.8579295           1
8       0   8  1285 -288   1852   -309   -356   916    150            -22  0.1283940           0
9       0  29    30   19    511      4      4   504      2             63  0.9954394           1
10      0   5   191   43    216    -10    -15    97     37             23  0.8866070           1
11      0   1   432   36    478     10     -1   156     13              8  0.6928087           1
14      0   2 11480  672    220    532   -120    22      7              6  0.6649320           1
```

Y su representación gráfica mediante los correspondientes histogramas:

Ilustración 44: Variable Estado

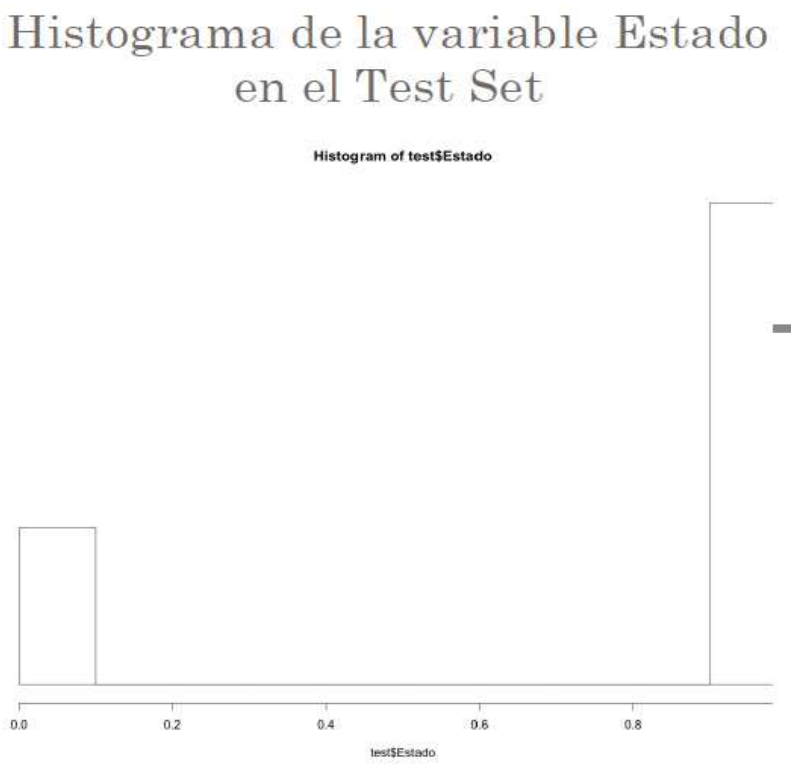


Ilustración 45: Predicción Estado

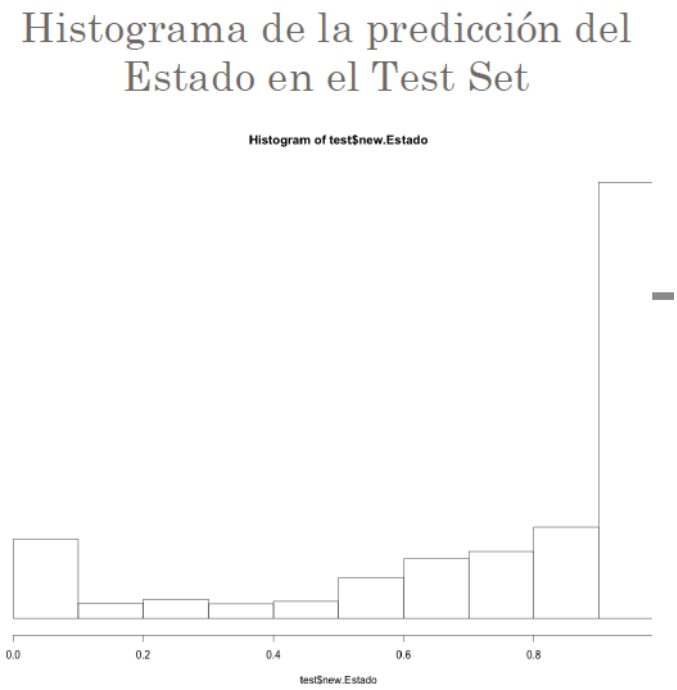
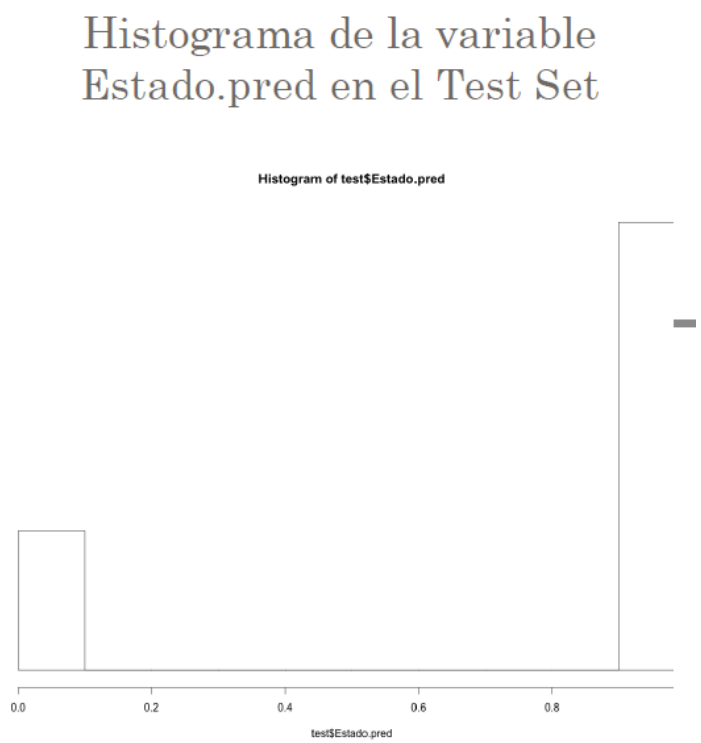


Ilustración 46: Variable Estado.pred



Como ejemplo de varias de las simulaciones tenemos:

Tabla 30: Simulación 1 Modelo de Regresión Logística

```
Start:  AIC=398014
Estado ~ Tes + TA + FP + Ventas + ResExp + ResEje + Gpers + Gamort +
    FonProp.TotAct

                Df Deviance    AIC
- Tes            1     7128   7146
- ResEje         1     7132   7150
- FP             1     7136   7154
- TA             1     7157   7175
- ResExp         1     7178   7196
- Ventas         1   212874 212892
- Gamort         1   366204 366222
<none>               397994 398014
- Gpers          1   490266 490284
- FonProp.TotAct 1   663996 664014
```

Tabla 31: Simulación 2 Modelo de Regresión Logística

```
Step:  AIC=7146.18
Estado ~ TA + FP + Ventas + ResExp + ResEje + Gpers + Gamort +
    FonProp.TotAct

                Df Deviance     AIC
- Gamort         1   7128.8  7144.8
- Gpers          1   7129.4  7145.4
<none>               7128.2  7146.2
- Ventas         1   7130.4  7146.4
- ResEje         1   7135.6  7151.6
- FP             1   7138.4  7154.4
- TA             1   7158.6  7174.6
- ResExp         1   7185.0  7201.0
- FonProp.TotAct 1  12689.7 12705.7
```

Tabla 32: Simulación 3 Modelo de Regresión Logística

```
Step:  AIC=7144.84
Estado ~ TA + FP + Ventas + ResExp + ResEje + Gpers + FonProp.TotAct

                Df Deviance     AIC
- Gpers          1   7129.9  7143.9
- Ventas         1   7130.7  7144.7
<none>               7128.8  7144.8
- ResEje         1   7137.1  7151.1
- FP             1   7138.4  7152.4
- TA             1   7158.6  7172.6
- ResExp         1   7189.6  7203.6
- FonProp.TotAct 1  12725.4 12739.4
```

Tabla 33: Simulación 4 Modelo de Regresión Logística

```
Step:  AIC=7143.87
Estado ~ TA + FP + Ventas + ResExp + ResEje + FonProp.TotAct

                Df Deviance    AIC
<none>               7130   7144
- Ventas         1     7137   7149
- ResEje         1     7138   7150
- FP             1     7138   7150
- TA             1     7159   7171
- ResExp         1     7190   7202
- FonProp.TotAct 1   226498 226510
```

Tabla 34: Resumen Simulación Modelo de Regresión Logística

```
CrossTable(test$Estado,test$Estado.pred)
Total Observations in Table:  3280

             | test$Estado.pred
 test$Estado |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |       607 |       155 |       762 |
             |   977.093 |   310.002 |           |
             |     0.797 |     0.203 |     0.232 |
             |     0.768 |     0.062 |           |
             |     0.185 |     0.047 |           |
-------------|-----------|-----------|-----------|
           1 |       183 |      2335 |      2518 |
             |   295.689 |    93.813 |           |
             |     0.073 |     0.927 |     0.768 |
             |     0.232 |     0.938 |           |
             |     0.056 |     0.712 |           |
-------------|-----------|-----------|-----------|
Column Total |       790 |      2490 |      3280 |
             |     0.241 |     0.759 |           |
-------------|-----------|-----------|-----------|
```

```
> aciertos <-
sum(test$Estado*test$Estado.pred) +
sum((!test$Estado)*(!test$Estado.pred))

> prop.aciertos <- aciertos/nrow(test)

> prop.aciertos
[1] 0.8969512
```

Ilustración 47: Resumen Simulación Modelo de Regresión Logística

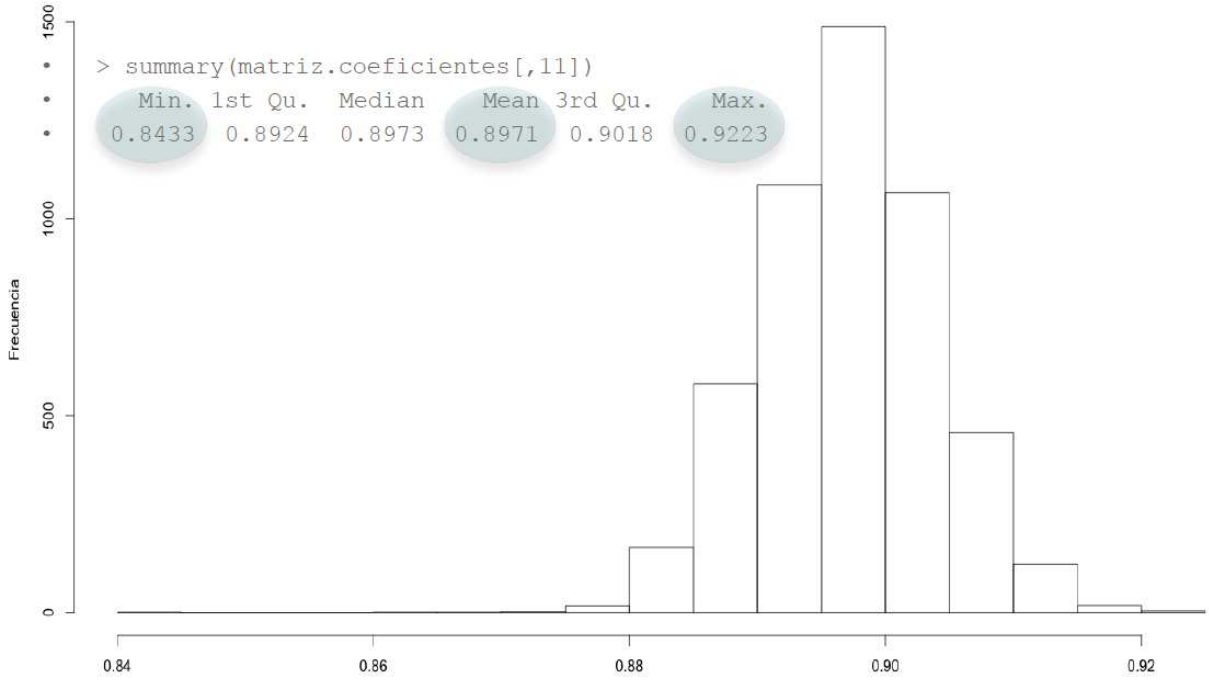
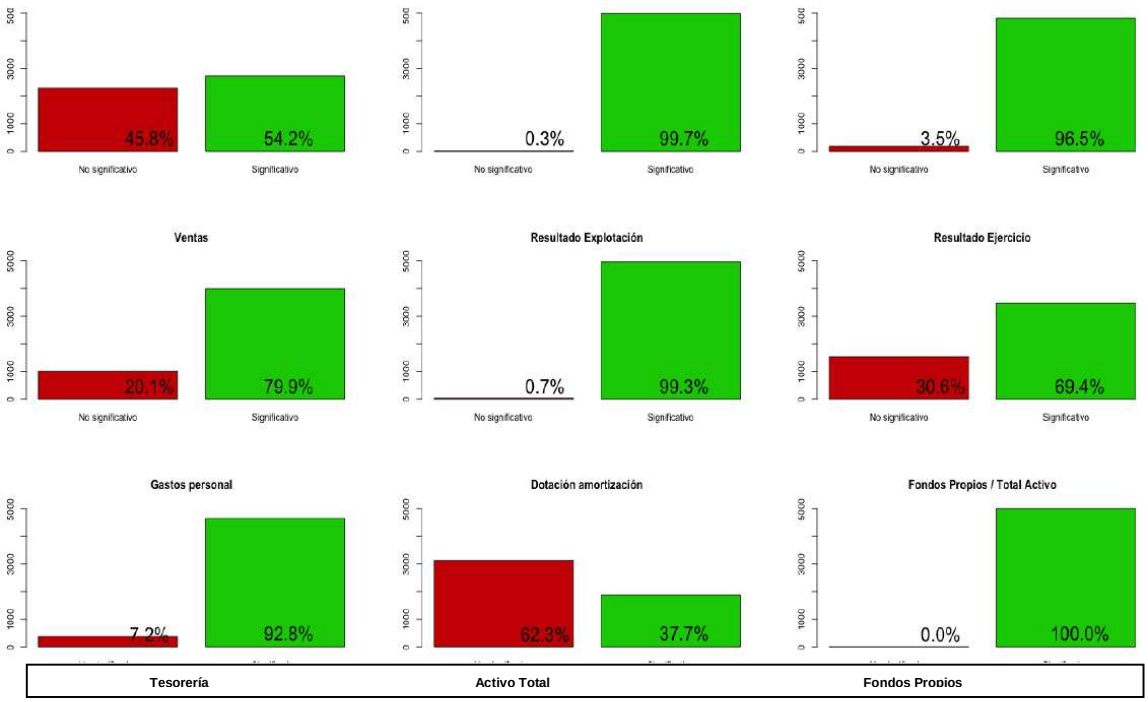


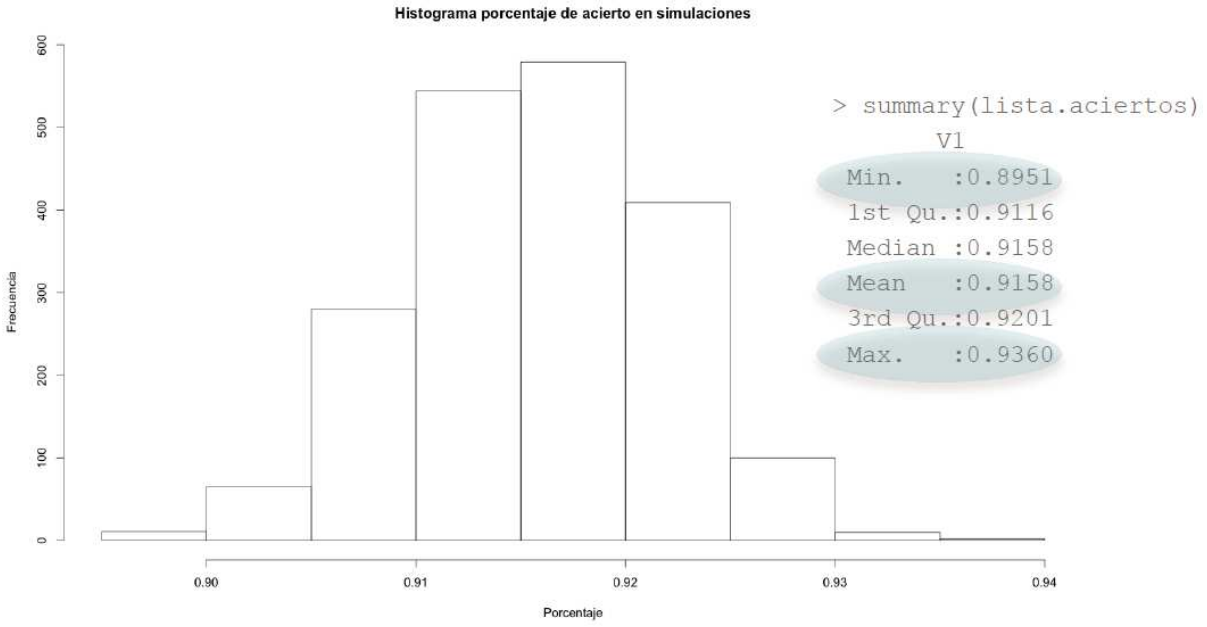
Ilustración 48: Significatividad de las variables.



6.2. Modelo SVM

Se han llevado a cabo 2000 simulaciones en este caso por ser el tiempo de computación muy elevado.

Ilustración 49: Resultado Simulación Modelo SVM



6.3. Modelo de particionado recursivo

Ilustración 50: Resultado Simulación Modelo Particionado Recursivo

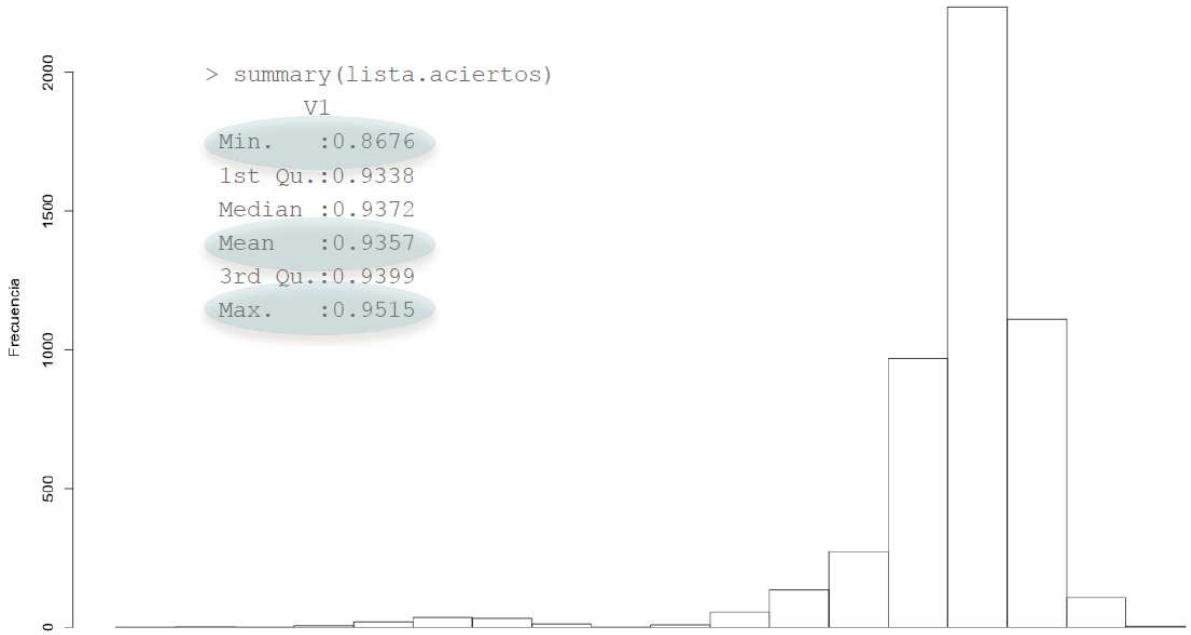


Ilustración 51: Ejemplo de estructura de nodos en una simulación concreta

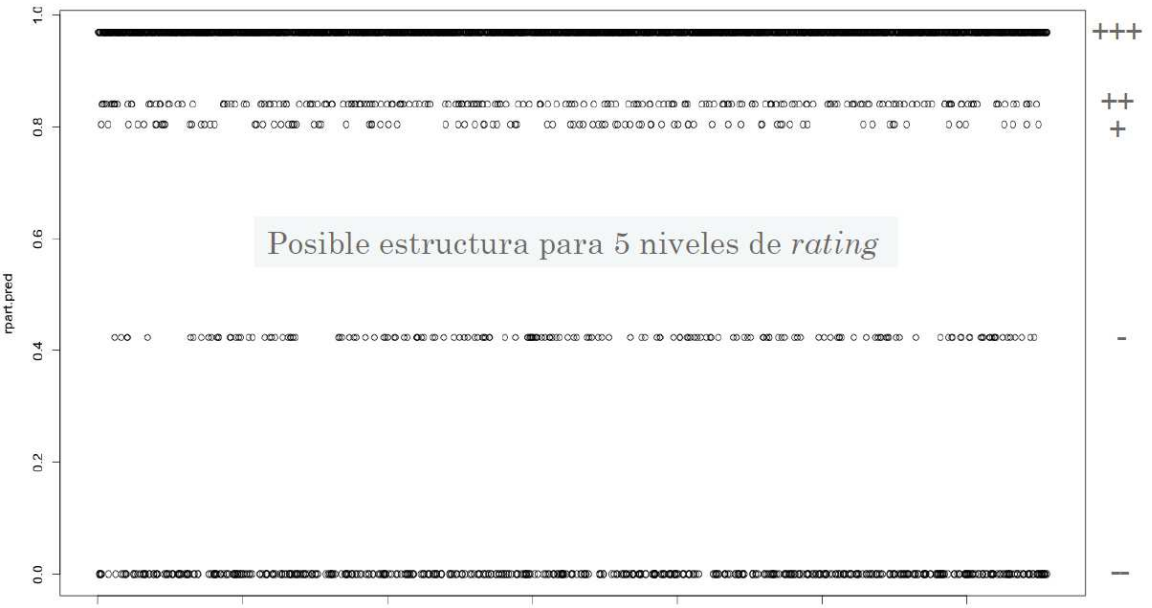


Tabla 35: Ejemplo simulación Modelo Particionado Recursivo

```
> table(rpart.pred)
rpart.pred
               0 0.423135464231355 0.803977272727273 0.840182648401826 0.968645391653649
             561               170                96               216              2236

> table(testset$Estado)
   0    1
 780 2499

> sum(testset[rpart.pred==names(table(rpart.pred)[1]),c("Estado")]==0)
[1] 561
> sum(testset[rpart.pred==names(table(rpart.pred)[1]),c("Estado")]==1)
[1] 0

> sum(testset[rpart.pred==names(table(rpart.pred)[2]),c("Estado")]==0)
[1] 89
> sum(testset[rpart.pred==names(table(rpart.pred)[2]),c("Estado")]==1)
[1] 81


> sum(testset[rpart.pred==names(table(rpart.pred)[3]),c("Estado")]==0)
[1] 22
> sum(testset[rpart.pred==names(table(rpart.pred)[3]),c("Estado")]==1)
[1] 74

> sum(testset[rpart.pred==names(table(rpart.pred)[4]),c("Estado")]==0)
[1] 36
> sum(testset[rpart.pred==names(table(rpart.pred)[4]),c("Estado")]==1)
[1] 180

> sum(testset[rpart.pred==names(table(rpart.pred)[5]),c("Estado")]==0)
[1] 72
> sum(testset[rpart.pred==names(table(rpart.pred)[5]),c("Estado")]==1)
[1] 2164


> print(rpart.model)
n= 13119

node), split, n, deviance, yval
      * denotes terminal node

 1) root 13119 2388.97000 0.7605763
   2) FP< 0.5 2272    0.00000 0.0000000 *
   3) FP>=0.5 10847  799.38070 0.9198857
     6) ResEje< -83.5 1009  249.08620 0.5559960
      12) FonProp.TotAct< 34.5 657  160.36830 0.4231355 *
      13) FonProp.TotAct>=34.5 352   55.47443 0.8039773 *
     7) ResEje>=-83.5 9838  402.98400 0.9572067
      14) ResEje< -25.5 876  117.62560 0.8401826 *
      15) ResEje>=-25.5 8962  272.18940 0.9686454 *
```

Ilustración 52: Ejemplo simulación Modelo Particionado Recursivo

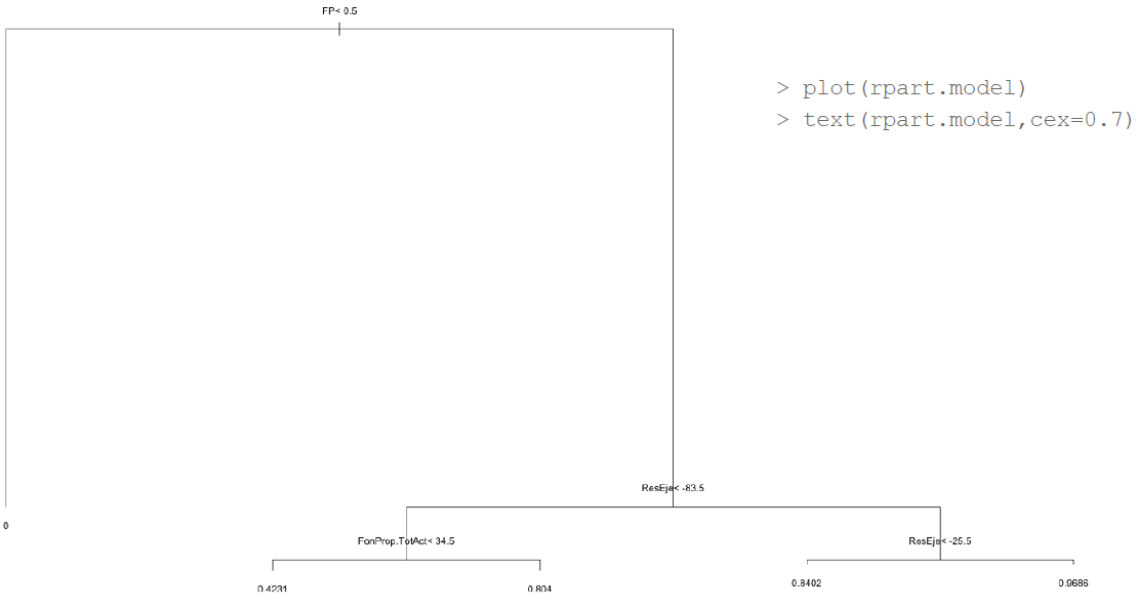


Tabla 36: Resultados Simulación Modelo Particionado Recursivo

Total Observations in Table: 16398

datos$Estado	datos$Estado.pred 0	1	Row Total
0	3358	563	3921
	6539.453	1986.742	
	0.856	0.144	0.239
	0.879	0.045	
	0.205	0.034	
1	463	12014	12477
	2055.077	624.350	
	0.037	0.963	0.761
	0.121	0.955	
	0.028	0.733	
Column Total	3821	12577	16398
	0.233	0.767	

```
> prop.aciertos
[1] 0.9374314
```

7. Conclusiones

A continuación se exponen las principales conclusiones que se han obtenido del presente Trabajo Final de Máster, comenzando con las que se han ido obteniendo en cada uno de los capítulos y terminando con unas conclusiones finales, en las que se expone la calidad acerca de los resultados obtenidos y se expone el grado de cumplimiento de los objetivos que se han fijado al inicio del Trabajo Final de Máster.

También se destacan las principales aportaciones que se han realizado en el Trabajo Final de Máster y se concluye con las futuras líneas de investigación que se pretende desarrollar.

En el capítulo introductorio se ha destacado necesidad e importancia de tener una metodología adecuada para la predicción de la insolvencia empresarial de cara a los posibles riesgos crediticios que la situación de estas puede generar.

Tras una revisión del estado del arte se ha comprobado que no hay estudios similares de utilización de la técnica de construcción de árboles de decisión para la predicción de la insolvencia empresarial en muestras semejantes a la analizada y mediante el algoritmo empleado.

El riesgo de crédito ha pasado a tener una importancia crucial y de la cual destaca entre otras, la reforma de la regulación financiera internacional, iniciada, para las entidades bancarias, con la normativa establecida en el Nuevo Acuerdo de Capital de Basilea -Basilea II-, de reciente implantación, en la que la medición y gestión de los riesgos financieros a los que están expuestas dichas entidades tiene una trascendental importancia.

Para el cálculo de la probabilidad de incumplimiento o de insolvencia, como se ha visto en los distintos apartados, existen diferentes metodologías estadísticas con las que es posible predecir la probabilidad de insolvencia o fallo en un periodo dado, entre las que destacan los modelos los métodos basados en el factor humano (sólo la decisión del humano y el factor humano más una normas), los modelos estadísticos (univariantes,

multivariantes y de regresión), y finalmente los modelos de aprendizaje de máquinas (modelos de inteligencia artificial, redes neuronales artificiales y modelos de árboles de decisión) utilizados a menudo en el sector financiero para realizar esta tarea.

En este TFM se presenta el uso de árboles de decisión como herramienta para el cálculo de probabilidades de insolvencia, por cuanto representa un método efectivo para la estimación, al igual que la mayoría de los métodos alternativos, pero ofrece la ventaja fundamental al ser un método de fácil entendimiento para personas que no cuentan con conocimientos avanzados de estadística. El modelo permite realizar una clasificación de clientes considerados como buenos y de máxima solvencia (probabilidades bajas de incumplimiento) y hasta clientes considerados como malos con altas probabilidades de insolvencia.

En el capítulo en el que se analiza el estado del arte, nos hemos centrado en los estudios que creemos más importantes relacionados con el campo de la insolvencia empresarial y en los que se analiza con detalle todos los pasos realizados por los autores para llegar a sus conclusiones, que han sido plasmados en estos estudios y aquí reflejados.

En el primer apartado de este capítulo se analiza la validez de la información contable como base óptima en la predicción de la insolvencia empresarial, así como la capacidad de los modelos univariantes para clasificar de forma acertada la posible insolvencia empresarial.

La muestra sobre la que se realiza el estudio son empresas españolas de gran dimensión sujetas a la CMNV, no pertenecientes al sector financiero ni de seguros. Sobre una muestra de más de 600 empresas se identifican 30 fracasadas a las que se les empareja frente a una que no ha fracasado del mismo sector, activo y año.

Los resultados indican que varios ratios contables consiguen clasificar las empresas en solventes e insolventes con un grado de acierto del 95% para el año anterior al fracaso. Los indicadores de *Rentabilidad del Activo, Margen de Beneficio del Resultado Ordinario y Cobertura de Gastos Financieros* fueron los que resultaron con mayor poder discriminante. Por lo anterior se concluye que modelos univariantes pueden

presentar expresiva capacidad predictiva del riesgo de insolvencia, además de contar con mayor practicidad en su proceso de estimación e implementación.

El estudio más clásico, basado en ratios contables, es el de Beaver (1966) que, a pesar de su antigüedad, conserva todavía su vigencia metodológica. El objetivo del estudio de Beaver (1966) es la predicción de la insolvencia empresarial a través de ratios, concluyendo que el ratio Cash flow a Deuda total es el de mayor valor predictivo.

La esencia del estudio de Beaver es el análisis dicotómico, que es un test predictivo, cuya finalidad última es seleccionar que ratio permite una mejor discriminación entre ambos grupos de empresas. El punto de corte se determina mediante un proceso de prueba y error, por el cual se van fijando diferentes valores y se va tanteando cuál de ellos produce menores errores de clasificación, de tal forma que si:

$Ratio_i \leq$ punto de corte = clasificación: fracasada

$Ratio_i >$ punto de corte = clasificación: sana

Los Tipos de Error ante una clasificación dicotómica pueden ser:

Error Tipo I: clasificar una empresa fallida como sana

Error Tipo II: clasificar una empresa sana como fallida

Las consecuencias de incurrir en un tipo u otro de error son claramente distintas, así por ejemplo en el caso de concederse un préstamo o decidir invertir en una empresa, el coste de error Tipo I es mucho mayor que el Tipo II, puesto que el Tipo II es un coste de oportunidad asociado a la no elección de dicha empresa; en cuanto el Tipo I involucra la pérdida de parte o totalidad del capital invertido.

Del análisis de perfiles de los ratios se concluye que los ratios de liquidez que muestran las empresas fracasadas son menores que los de las sanas empeorando levemente cuando se acerca la fecha de fracaso. , los ratios de rentabilidad revelan ser menores en las empresas fracasadas con una caída bastante acentuada de ésta al acercarse la fecha de fracaso, los ratios de rotación exhiben menor rotación en las empresas

fracasadas, finalmente los ratios de endeudamiento, denotan en todos ellos mayores niveles de endeudamiento en las empresas fracasadas que en las sanas.

Los ratios R19 (RA/AT: Resultado actividad/Activo Total) y R70(RE/GF: Resultado explotación/Gastos financieros) son los que proporcionan mejores resultados a nivel univariante. Ambos arrojan un porcentaje de acierto global de 95% para el año inmediato anterior al fracaso y de 83% para el segundo año anterior. Y a fin de elegir el mejor ratio que discrimine a las empresas fracasadas y sanas, se seleccionó el ratio R19 como la mejor variable univariante, ya que arroja un porcentaje de error Tipo I menor que el R70 y, como fue comentado previamente, se considera este error más grave que el Tipo II.

En el segundo apartado se realiza un estudio de la predicción de la insolvencia a partir de una muestra similar a la del anterior estudio, así como los mismos ratios y variables definitorias de la situación financiero-contable de las empresas, enfrentando una insolvente frente a una solvente, de las mismas características y comparables.

En este caso los métodos utilizados han sido el análisis discriminante y el análisis logit. Ambos son modelos estadísticos multivariantes, que ya han sido explicados en capítulos introductorios y que también muestran altos porcentajes de acierto en la clasificación de las empresas, siendo éstos en torno del 95% para el año anterior al fracaso y del 80% en el segundo año anterior.

Los resultados de la aplicación del modelo de análisis discriminante nos proporcionan una función para el primer año anterior al fracaso que la componen los ratios: R20, R36 y R56, los cuales fueron determinados en tres etapas, habiéndose incorporado en ese orden.

$$Z = 0{,}583 + 7{,}937 \cdot R20 + 0{,}352 \cdot R36 - 2{,}403 \cdot R56$$

Esta función resultante alcanza un porcentaje de acierto global de 93,3%.

El ratio R20 de rentabilidad, RAIT y Ext./AT, definido como resultado antes de intereses, impuestos y extraordinarios a activo total, es el que más contribuye en la discriminación de los grupos. El ratio R36, también de rentabilidad, CFT/RE,

determinado por la relación del cash flow tradicional a resultados de la explotación, fue seleccionado en la segunda etapa del proceso. El ratio R56 de endeudamiento, ELP/AT, definido como exigible a largo plazo a activo.

La función discriminante del modelo para el segundo año anterior al fracaso quedó constituida de la siguiente manera:

$$Z = 0{,}213 + 14{,}672 \cdot R20$$

Esta función resultante alcanza un porcentaje de acierto global del 80%.

En la estimación de los modelos multivariantes a través del análisis logit, los ratios que finalmente entran en el modelo para el primer año anterior al fracaso son R06, R20, R23 y Constante, con el que se alcanza un porcentaje de éxito en la clasificación del 95%. Los resultados alcanzados dos años antes del fracaso derivado a través del método por pasos denotan la incorporación de un único ratio, el R20 (resultado antes de intereses, impuestos y extraordinarios a activo total), como variable independiente, además de la constante. Los resultados del modelo propuesto para el segundo año anterior al fracaso, reflejan un porcentaje de acierto global de 81,4%.

En el tercer apartado se realiza un estudio mediante sistemas de inducción de reglas y árboles de decisión, pero mediante el programa comercial See5, que es la versión extensa del algoritmo ID3 y C4.5, que posteriormente se expondrá un ejemplo, desarrollado por Quinlan en 1997.

El algoritmo See5 permite construir automáticamente a partir de un conjunto de datos de entrenamiento un árbol de clasificación. Para inferir el árbol, el algoritmo realiza particiones binarias sucesivas en el espacio de las variables explicativas, de forma que en cada partición se escoge la variable que aporta más información en función de una medida de entropía o cantidad de información. El árbol así construido consta del mínimo número de atributos (variables) que se requieren para la clasificación correcta de los ejemplos dados, con lo que es claro el alto poder explicativo de esta técnica.

También se pueden elaborar, a partir del árbol, reglas de clasificación fácilmente interpretables, que definen las características que más diferencian a las distintas

categorías establecidas inicialmente. Este tipo de sistemas clasificadores presentan la ventaja, frente a las técnicas estadísticas, de que tienen un carácter estrictamente no paramétrico. Además, aunque no alcanzan el poder predictivo de las redes neuronales, sus resultados son mucho más fácilmente interpretables.

En este caso la muestra se realiza para un conjunto de empresas del sector seguros no-vida, también de suma importancia el predecir su insolvencia temprana por el riesgo que conlleva este campo. Consta de 36 empresas no fracasadas y 36 empresas fracasadas, emparejadas por tamaño y tipo de negocio. El éxito o fracaso de una empresa será entendido como una variable dependiente que deberá ser explicada por un conjunto de ratios financieros que actuarán como variables independientes. Los ratios coinciden en su mayoría con los utilizados en los anteriores estudios, además de añadir unos ratios específicos del sector de los seguros. De las 72 empresas de que consta la muestra original, se han utilizado únicamente 27 empresas de cada una de las submuestras para la elaboración de los modelos, reservando las 9 restantes para poder comprobar la validez de los mismos. En consecuencia, se tiene una muestra de entrenamiento para obtener los árboles y reglas de decisión formada por 54 empresas y una muestra de validación para verificar su capacidad predictiva formada por 18 empresas.

En el árbol aparecen únicamente 6 de los 19 ratios iniciales, lo que indica que 13 de los ratios empleados no aportan información relevante para clasificar las empresas como “buenas” o “malas”. El árbol proporciona el menor número de ratios necesarios para alcanzar el objetivo deseado. La evaluación de este árbol de decisión construido con la muestra de entrenamiento (54 empresas) indica que el árbol consta de 8 ramas y comete un total de 3 errores (5,6%), lo que supone un porcentaje de aciertos del 94,4%; de la muestra de validación se obtiene un porcentaje de clasificaciones correctas del 72,2%. El algoritmo See5 incorpora un método llamado *adaptive boosting*, basado en el trabajo de Freund y Schapire (1997), que en generar varios clasificadores (árboles o conjuntos de reglas) en vez de sólo uno. Como primer paso, se construye un único árbol (o conjunto de reglas), que cometerá algunos errores en la clasificación. Estos errores serán el foco de atención al construir el segundo clasificador en aras de

corregirlos. En consecuencia, el segundo clasificador generalmente será diferente al primero y también cometerá errores que serán el foco de atención durante la construcción del tercer clasificador. Este proceso continúa para un número predeterminado de iteraciones o trials. Mediante este procedimiento, se consigue obtener un clasificador verdaderamente preciso. Así, partiendo del primer árbol de decisión los resultados que se alcanzan realizando 18 iteraciones son con la muestra de entrenamiento, el 100% de clasificaciones correctas y con la muestra de validación, el 83,3% de clasificaciones correctas:

Este método viene a asemejarse al método que después aplicaremos en nuestro algoritmo de Particionado Recursivo, en esencia es clasificar la muestra en varios clasificadores que mejoraran el resultado predictivo del algoritmo.

En el cuarto caso el trabajo de investigación desarrollado se asemeja a nuestro estudio por cuanto que se utiliza el algoritmo Part y su comparación con el método estadístico multivariante de la Regresión Logística, pero aplicado al sector de seguros. no vida.

La esencia de este algoritmo de inducción de árboles de decisión es la siguiente, por una parte se basan en el enfoque -divide and conquer- y paralelamente lo hacen con la estrategia - separate and conquer -, que coincide con el particionado.

La muestra sobre la que se aplica este modelo coincide con la utilizada en el anterior compuesta por un total de 72 empresas y una muestra de ratios semblante a la del anterior estudio.

Aquí se estima que la separación de la muestra dejando una parte de ella para la validación, supone un empeoramiento de la calidad de los modelos, con lo que se sigue el proceso de validación - jackknife (Efron, 1982)-, también denominado - leave one out- Siendo k el número de instancias que contenga el conjunto de entrenamiento, 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 este, ofrece la oportunidad de conseguir la estimación más exacta que posiblemente pueda obtenerse.

Para la aplicación del algoritmo PART y la Regresión Logística a esta muestra se ha utilizado en este estudio, el paquete gratuito de minería de datos WEKA desarrollado en la Universidad de Waikato (Witten y Frank, 2000) y el software R 2.1.0 distribuido gratuitamente por CRAN Foundation (R Development Core Team, 2005).

En cuanto a los resultados que ofrece el algoritmo, para el año anterior a la insolvencia tenemos tres reglas, la primera si se cumplen las condiciones R3 (activo circulante/pasivo circulante) mayor 1.186837 y R14 (Gastos técnicos seguro directo/(fondos propios + provisiones técnicas) menor o igual que 0.67952 la empresa será calificada como sana. La segunda si el R3 es menor o igual a 1.974321 la empresa será clasificada como fracasada y finalmente por defecto si no se clasifica con la segunda vendrá clasificada como fracasada. El porcentaje de acierto en la estimación se acerca al 82%. Este modelo confirma la importancia de la liquidez de cara a predecir el fracaso empresarial, medida a través del ratio R3. Aunque la liquidez es una necesidad generalizada en cualquier tipo de empresa, en la empresa aseguradora dicha necesidad reviste una mayor importancia.

Con los datos del segundo año previo a la quiebra, se obtiene la siguiente lista de decisión de dos reglas. la primera si se cumplen las condiciones R3 mayor 0.912177 y R14 menor o igual que 0.616583, la segunda y por defecto si no se clasifica con la primera vendrá clasificada como fracasada. El porcentaje de acierto en la estimación se acerca al 72%.

Para el modelo de Regresión lineal también se aplica el método – jackknife- siendo el porcentaje de acierto en la clasificación del 68% para el primer año anterior a la insolvencia y del 70,5% para el segundo año anterior, claramente inferiores a los obtenidos mediante el algoritmo PART, que además proporciona modelos de interpretación más sencilla y a su vez es más robusto ante el "ruido" introducido por los valores faltante o "outliers", y por tanto se adecua mejor a la información contable que a menudo presenta datos interrelacionados, incompletos, adulterados o erróneos.

En el apartado cinco del presente estudio, nos centramos en el desarrollo y explicación del propio trabajo, la base de datos utilizada , el lenguaje de programación en el que se han implementado los modelos y algoritmos utilizados de forma detallada.

Se explica en este apartado la importancia de la minería de datos, a la que podríamos bautizar como la ciencia que explora y analiza las bases de datos, cada día más cuantiosas y complejas, para la extracción e identificación no trivial de patrones.(Conocimiento).

La base de datos que se ha utilizado en el presente estudio es la base de datos SABI, -Sistema de Análisis de Balances Ibéricos- . Posee un software avanzado que permite conocer y analizar los balances de más de 1,4 millones de empresas españolas y más de 400.000 portuguesas. Permite análisis detallados, estadísticos y comparativos de empresas y grupos de empresas, así como la obtención de gráficos ilustrativos de los balances y cuentas de resultados. Ello facilita el seguimiento de la evolución financiera de las empresas en relación a sus competidores, así como los análisis del entorno de mercado/competencia y la investigación económica en general.

A continuación se centra el apartado del desarrollo en la elaboración e implementación de forma manual mediante el auxilio de la hoja de cálculo Excel del algoritmo raíz o seminal de los árboles de decisión el ID3, desarrollado por Quinlan en 1983.

Este algoritmo se basa en la teoría de la información, desarrollada en 1948 por Claude Elwood Shannon. Su significado es "inducción mediante árboles de decisión", capaz de tomar decisiones con gran precisión.

Es un sistema de aprendizaje supervisado que aplica la estrategia “divide y vencerás” para hacer la clasificación, implementando métodos y técnicas para la realización de procesos inteligentes, representando así el conocimiento y el aprendizaje, con el propósito de automatizar tareas.

El ID3 permite determinar el árbol de decisión mínimo, para un conjunto de objetos. Permite que la información se mantenga organizada y entendible para cualquier persona, además haciendo uso de una secuencia de preguntas, donde cada pregunta es evaluada con el propósito de obtener la mejor respuesta posible.

La idea básica del ID3 es de determinar, para un conjunto de ejemplos dados, el atributo más importante, es decir, aquel que posea mayor poder discriminatorio para dicho conjunto; este atributo es usado para la clasificación de la lista de objetos, basados en los valores asociados por él mismo. Después de haber hecho la primera prueba de atributo, esta arrojará un resultado, el cual es en sí mismo un nuevo problema de aprendizaje de árbol de decisión con la diferencia de que contará con menos ejemplos y un atributo menos, por lo que cada atributo que se selecciona se descarta para la siguiente prueba.

Sobre un ejemplo de 14 empresas y elegidas una serie de variables que se consideran más discriminantes por acortar el cálculo y entre los cuales se van entremezclando los casos, el objetivo es la aplicación manual y se forma sencilla del algoritmo ID3 y comprobar su funcionamiento. Mediante la construcción de la hoja de cálculo y la aplicación de la formulación del algoritmo tenemos la variable *fondos propios* como el atributo más importante, con el mayor poder discriminatorio para este conjunto de casos y este a su vez ramifica con el bajo (nulo) con el *endeudamiento* o alto con la *rentabilidad*, para a su vez ramificarse de nuevo de una forma sencilla y llegar a la predicción acertada de todos los casos.

Con muestras mayores y con mayor número de variables, ya sería imprescindible su programación y aplicación de algoritmos más avanzados como el que se utiliza posteriormente, la librería RPART implementado en el lenguaje de programación R.

Pero como adelanto y entreno ha servido para ver las grandes posibilidades que este algoritmo ofrece para cualquier rama de la ciencia.

Para finalizar se han desarrollado con mayor detalle la teoría de la Regresión Logística que luego se implementará en R con el auxilio de la función *glm*, también el modelo de máquina de vector soporte con el auxilio de los paquetes *kvm, svmlight y el e1071*, y finalmente el modelo de particionado recursivo, ya explicado anteriormente y que se implementará en R con ayuda del paquete *RPart.*

De los resultados obtenidos se puede destacar que mediante la regresión logística obtenemos un porcentaje de aciertos en la insolvencia del 89’70%, mediante el modelo SVM se obtiene el 91’58% de acierto y finalmente mediante el modelo de Particionado Recursiva, se alcanza el mayor porcentaje de acierto con un 93’74%.

Se concluye que tal y como se había comprobado en el estado del arte, los modelos de inducción mediante árboles de decisión son una herramienta muy válida y a tener en cuenta para la predicción de la insolvencia empresarial y el riesgo crediticio que esta situación de insolvencia empresarial va a provocar en la institución financiera .

8. Bibliografía

Guijarro, F (2013). Executive MBA. *Modelos de predicción de insolvencia empresarial y sistema de rating empresarial a partir de información económico-financiera.*

Hernández, P. A. C. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. *Revista colombiana de estadística*, *27*(2), 139-151.

Colauto, R. D., Pinheiro, L. E. T., & Pinheiro, J. L. (2009). Información Contable en la Predicción de Insolvencia: estudio inferencial univariante aplicado a empresas españolas. *Revista Contemporânea de Contabilidade, UFSC, Florianópolis*, *1*(12), 151-170.

Collazos, J. A. A., Castaño, A. H. D., & Ocampo, E. M. T. (2012). Una comparación entre métodos estadísticos clásicos y técnicas metaheurísticas en el modelamiento estadístico. *Scientia Et Technica*, *17*(50), 68-77.

Colauto, R. D., Pinheiro, L. E. T., & Pinheiro, J. L. (2009). Información Contable en la Predicción de Insolvencia: estudio inferencial univariante aplicado a empresas españolas. *Revista Contemporânea de Contabilidade, UFSC, Florianópolis*, *1*(12), 151-170.

Colauto, R. D., Pinheiro, L. E. T., & Pinheiro, J. L. (2009). Información Contable en la Predicción de Insolvencia: estudio inferencial univariante aplicado a empresas españolas. *Revista Contemporânea de Contabilidade, UFSC, Florianópolis*, *1*(12), 151-170.

Pinheiro, L. E. T., & Pinheiro, J. L. (2009). Modelos de Evaluación del Riesgo de Insolvencia de Empresas Españolas Cotizadas. *Contabilidade Vista & Revista*, *19*(3), 95-121.

Díaz Martínez, Z., Fernández Menéndez, J., & Segovia Vargas, M. J. (2004). Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

Díaz Martínez, Z., Fernández Menéndez, J., & Pozo García, E. M. (2006). La inteligencia artificial como una alternativa viable en la predicción de la insolvencia de empresas de seguros. *Análisis financiero*, (100), 64-75.

Fana, J. A. G., Martínez, A. H., Zanón, J. L. V., & Arellano, A. S. (2003).*Predicción de insolvencias con el método Rough Set*. Universidad Complutense de Madrid, Facultad de Ciencias Económicas y Empresariales.

Vargas, M. S., Fana, J. G., Martínez, A. H., Zanón, J. V., & Contabilidad, I. La metodología Rough Set frente al Análisis Discriminante en los problemas de clasificación multiatributo.

Segovia Vargas, M. J., Gil Fana, J. A., Heras Martínez, A., & Vilar Zanón, J. L. (2003). Predicción de insolvencias con el método Rough Set.

Moreno B. Minería Sobre Grandes Cantidades de Datos.

Sempere, J. Aprendizaje de árboles de decisión. *Universidad Politécnica de Valencia, Valencia*.

Trigo Martínez, E. (2009). Análisis y medición del riesgo de crédito en carteras de activos financieros ilíquidos emitidos por empresas.

Varguez-Moo, M., Uc-Cetina, V., & Brito-Loeza, C. (2014). Clasificación de documentos usando Máquinas de Vectores de Apoyo. *Abstraction and Application Magazine*, *6*.

Therneau, T. M., & Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines.

Menéndez-Plans, C., Orgaz, N., & Prior, D. (2012). ¿ Existe relación entre la información contable y el riesgo sistemático de las empresas? Estimación con datos de panel. *Academia. Revista Latinoamericana de Administración*, (49), 1-16.

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

Sanchís Arellano, A. (2003). *Una aplicación del análisis discriminante a la previsión de la insolvencia en las empresas españolas de seguros no-vida*. Universidad Complutense de Madrid, Servicio de Publicaciones.

Servente, M., & García, R. (2002). Algoritmos TDIDT aplicados a la minería de datos inteligente. *Universidad de Buenos Aires, Buenos Aires*.

Daza, D. P. (2014). Sistema de inteligencia embebida con autoaprendizaje basado en una arquitectura de árbol de decisión dinámico y adaptativo.

Guevara Maldonado, C. B. (2011). Reconocimiento de patrones para identificación de usuarios en accesos informáticos.

Martínez, Z. D., & Menéndez, J. F. (2006). El Algoritmo See5 versus la metodología Rough Set. Una aplicación a la predicción de la insolvencia en empresas Españolas de seguros no-vida. *Cuadernos de Estudios Empresariales*, (15), 179-198.

Díaz Martínez, Z., Fernández Menéndez, J., & Segovia Vargas, M. J. (2004). Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras.

Segovia, M. J., Gil, J. A., VILAR, L., & HERAS, A. (2003). La metodología rough set frente al análisis discriminante en la predicción de insolvencias en empresas aseguradoras. In *Anales del Instituto de Actuarios Españoles* (Vol. 9, pp. 153-180).

Silva, J. O. D., Wienhage, P., Souza, R. P. S. D., Bezerra, F. A., & Lyra, R. L. W. C. D. (2012). Capacidade Preditiva de Modelos de Insolvência com Base em Números Contábeis e Dados Descritivos. *Revista de Educação e Pesquisa em Contabilidade (REPeC)*, *6*(3).

González, V. M. (2009). Estructura de vencimiento de la deuda y riesgo de crédito en las empresas españolas. *Universia Business Review*, (22), 88-101.

NOGUER, B. G. D. A., & Muñoz, M. I. (2007). La calidad de los ajustes por devengo no afecta al coste de la deuda de las pymes españolas.*Investigaciones económicas*, *31*(1), 79-117.

Molina, M. J. C. UTILIDAD DE LA INFORMACIÓN ECONÓMICO-FINANCIERA PARA DECISIONES DE RIESGO CREDITICIO: ANÁLISIS DE PRÉSTAMOS CORPORATIVOS.

Pra, I., Ríos, A., ARGUEDAS, R., & Casals, J. (2010). *Gestión y control del riesgo de crédito con modelos avanzados*. Ediciones Académicas.

Trigo Martínez, E. (2009). Análisis y medición del riesgo de crédito en carteras de activos financieros ilíquidos emitidos por empresas.

Maldonado, S., & Weber, R. (2012). Modelos de Selección de Atributos para Support Vector Machines. *Revista Ingeniería de Sistemas*, *26*.

Bonillo, M. L. (2003). *Razonamiento Basado en Casos aplicado a Problemas de Clasificación*. Ph. D. thesis, Universidad de Granada.

Rodríguez Enríquez, E. (2003). La decisión de dividendos en las sociedades asturianas. Una aplicación del algoritmo See5 de Quinlan. *RAE: Revista Asturiana de Economía, 26*.

Daza, D. P. (2014). Sistema de inteligencia embebida con autoaprendizaje basado en una arquitectura de árbol de decisión dinámico y adaptativo.

Guevara Maldonado, C. B. (2011). Reconocimiento de patrones para identificación de usuarios en accesos informáticos.