ÁRBOLES DE CLASIFICACIÓN: UNA METODOLOGÍA PARA EL ANÁLISIS DE CRISIS BANCARIAS.

María José Vázquez Cueto
Departamento de Economía Aplicada III
Universidad de Sevilla
e-mail: pepi@us.es
Dolores Gómez Domínguez
Departamento de Economía Aplicada III
Universidad de Sevilla
e-mail: dogomez@us.es

Resumen

Los árboles de clasificación son una alternativa metodológica a los métodos multivariantes de la estadística clásica (modelo discriminante, logit, probit,...) cuando las variables implicadas en los análisis no verifican las hipótesis de partida de dichos métodos. En este trabajo comparamos el poder clasificatorio del árbol construido bajo determinadas especificaciones con el que se obtiene aplicando el análisis logit, en el estudio de los determinantes de las crisis bancarias ocurridas en la última década del siglo pasado. Para una muestra de 40 países, que incluye tanto economías industrializadas como economías en vías de desarrollo, y con los comportamientos de diez ratios, macroeconómicos y más específicos del sector financiero, durante el periodo 1988-2000, ambas metodologías señalan a las mismas variables como las más significativas. Sin embargo, los análisis muestran que el árbol construido obtiene menores porcentajes de clasificación erróneos en la muestra utilizada como validación, además de una más clara e intuitiva representación de los resultados obtenidos.

Palabras clave: Análisis logit, árboles de clasificación, ratios financieros, crisis bancarias.

Area temática: Métodos Cuantitativos.

Introducción

El análisis financiero utiliza con frecuencia el análisis logístico para tratar con los problemas de clasificación, aunque raramente discute sus limitaciones, que están bien documentadas en cualquier texto básico de estadística multivariante. En la mayoría de las investigaciones se supone tácita o explícitamente que las variables utilizadas como explicativas se distribuyen adecuadamente y que la muestra se ha elegido según las especificaciones del muestreo aleatorio simple. En la práctica, y aunque la no verificación de la hipótesis sobre la distribución de las variables sólo afecta marginalmente, la existencia de outliers, pequeños tamaños muestrales o grupos desproporcionados, afectan a los resultados de la clasificación, haciendo que el modelo se vuelva inestable y ofrezca pobres resultados en las muestras de validación.

Cuando el objetivo es, precisamente, una buena clasificación, deben utilizarse técnicas alternativas. Este trabajo considera, en concreto, una alternativa no paramétrica que se conoce como "partición recursiva binaria" o, más comúnmente, "árboles de clasificación".

Aunque los fundamentos teóricos de los árboles de clasificación se desarrollan en 1960, los requerimientos computacionales limitan sus aplicaciones hasta fechas muy recientes. Breiman et al (1984) fueron los responsables de introducirlos dentro de las técnicas estadísticas. Hoy en día existen varios Algoritmos que formulan árboles, SPSS, CART, C4.5, DTREG,... De entre todos hemos elegido el método denominado CART (Classification and Regression Trees). Como principales ventajas destacamos, entre otras, que: 1) No necesita hipótesis acerca de la distribución de las variables, 2) Puede trabajar con datos de distintos tipos: categóricos y continuos, 3) Sus resultados son robustos a los outliers, 4) Son invariantes a transformaciones monótonas de los datos, tales como el logaritmo neperiano, 5) Permite combinaciones lineales entre las variables, y 6) Selecciona automáticamente las variables que más reducen los errores de clasificación.

En este trabajo ponemos de manifiesto la utilidad de estos árboles cuando se aplican al análisis de las crisis bancarias. Para ello, y basándonos en una muestra de 40 países y en los valores que toman diez variables, que consideramos explicativas de la situación de crisis o no crisis bancaria, durante el periodo 1988-2000, realizamos un análisis logístico con el software SPSS 12 y construimos un árbol de clasificación con el algoritmo CART. La comparación de los resultados obtenidos, en términos de

porcentaje de clasificación correcta en la muestra original y en la muestra de validación, nos lleva a la consideración del mayor poder explicativo del árbol, además de su más fácil interpretación.

El trabajo se estructura de la siguiente forma. En el siguiente apartado se presenta la estructura de un árbol de clasificación y se exponen las ideas fundamentales del algoritmo CART. En el apartado tres se realizan las aplicaciones empíricas, cuyos resultados y conclusiones se exponen en el cuarto apartado.

Estructura de los árboles de clasificación binaria. Algoritmo CART

Un árbol de clasificación consta de tres tipos de nodos: raíz, interno y terminal. Existe un único nodo raíz que contiene a todas las observaciones. A partir de él se bifurcan dos ramas, cada una de las cuales da lugar a un nodo que puede ser interno o terminal. El nodo se dirá interno cuando, a su vez, se bifurca en dos ramas, en caso en que no se divida se dirá terminal. Cada nodo viene descrito por el subconjunto de la muestra que contiene. Este subconjunto, a su vez, viene descrito por intervalos de valores a los que pertenecen determinadas características o combinaciones lineales de las mismas. Así, si t es un nodo interno, se verá ramificado en dos nodos hijos, t_d y t_i, en base a una característica X o a una combinación lineal de características, $C(X_1, X_2,...,X_n)$, y un valor s. La característica X o la combinación de características $C(X_1, X_2,...,X_n)$ se selecciona de entre todas las existentes y el valor s se toma de tal forma que minimice la heterogeneidad de las dos submuestras resultantes.

Con el algoritmo CART la selección de las características que se incluirán en el árbol y la estructura del mismo es automática: En cada nodo busca el mejor valor de s para cada X y para cada posible combinación de características, y se queda con aquellos que producen el menor grado de diversidad.

La diversidad de un nodo está en relación con el valor de la función de impureza en el mismo (Breiman, 1988). Pueden definirse varias funciones de impureza, de entre ellas la más utilizada es la de Gini, definida de la siguiente forma: $g(t) = p_{1/t}\, p_{2/t}$, donde $p_{j/t}$ representa la proporción de casos que pertenecen a la clase j (j = 1,2) que ha sido asignada al nodo t.

La reducción del grado de diversidad viene medida por $g(t) - p_d\, g(t_d) - p_i\, g(t_i)$ donde p_k es la proporción de casos que van al nuevo nodo t_k (k = d,i). Los nodos siguen subdividiéndose mientras existan observaciones pertenecientes a varias clases y pueda reducirse el grado de diversidad.

Una vez que tengamos construido el árbol T con T´ nodos terminales, necesitamos una regla para asignar cada nodo terminal a una clase y una estructura de coste de clasificación errónea para evaluar los resultados del árbol.

Usualmente cada nodo terminal se asigna a la clase a la que mayoritariamente pertenecen sus elementos.

El coste esperado de clasificación errónea del árbol T se define como

$$R(T) = \sum_{j=1}^{2} \sum_{i=1, i\neq j}^{2} c(i/j)q(i/j)p(j)$$

donde c(i/j) es el coste de clasificar una observación de la clase j en la clase i, q(i/j) es la proporción de casos de la clase j clasificados erróneamente en la clase i y p(j) es probabilidad a priori de pertenencia a la clase j.

Este es un estimador mínimo del coste. Para mejorar la estimación del coste de clasificación erróneo del árbol, suele utilizarse el método de "validación cruzada". Dicho método selecciona aleatoriamente k submuestras de la muestra original y construye k árboles utilizando k-1 de la submuestras y validándolo para la que quedó fuera. De esta forma se obtienen k tasas de error, cuya media, $q^{cv}(i/j)$, reemplaza al valor de q(i/j) en R(T) permitiéndonos obtener la variable aleatoria $R^{cv}(T)$ que mejor estima el coste de clasificación errónea del árbol T.

La variable $R^{cv}(T)$ se utiliza también como criterio de parada en la poda del árbol máximo. Para ello se calcula el error estándar de la variable (SE(T)) y se va podando el árbol hasta conseguir un subárbol cuyo R^{cv} tenga una desviación estándar próxima a SE(T)[1].

Obviamente, dependiendo de las diferentes medidas de impureza, diferentes estructuras de costes, diferentes probabilidades a priori y varios niveles de SE, van obteniéndose distintos árboles. Un criterio que puede seguirse para elegir el mejor árbol es el de combinar la sensibilidad con la especificidad del mismo. La sensibilidad de un árbol hace referencia a la probabilidad estimada de clasificar un nuevo caso "malo"[2] como "malo", mientras que la especificidad consiste en la probabilidad contraria, es decir, clasificar un caso "bueno" como "bueno".

[1] Regla x SERULE, donde x toma valores entre 0 y 1 e indica la desviación permitida para el subárbol.

[2] Un caso se considera "malo" cuando representa una situación que no es deseable que ocurra.

Aplicación a la crisis bancarias

Para contrastar empíricamente la utilidad de los árboles en comparación con una técnica clásica de estadística multivariante como puede ser el análisis logit, procedemos a aplicar ambos procedimientos al análisis de las crisis bancarias ocurridas en la última década del siglo pasado. Para ello, y basándonos en una muestra de 40 países y en los valores que en ellos toman diez ratios durante el periodo 1988-2000, hemos construido, mediante el algoritmo CART un árbol de clasificación, y realizado un análisis logit con el software SPSS en su versión 12.

La tabla 1 recoge los 40 países utilizados en la muestra, diecinueve de ellos son economías industrializadas, mientras que los veintiuno restantes son economías en mayor o menor grado de desarrollo.

Tabla 1: Países considerados en la muestra.

INDUSTRIALIZADOS	EN DESARROLLO
ALEMANIA	BOLIVIA
AUSTRALIA	CHILE
BÉLGICA	CAMERÚN
CANADA	COSTA RICA
DINAMARCA	COLOMBIA
EEUU	COREA DEL SUR
ESPAÑA	ECUADOR
FINLANDIA	EGIPTO
FRANCIA	FILIPINAS
GRECIA	HONDURAS
HOLANDA	HUNGRÍA
ITALIA	INDIA
ISRAEL	INDONESIA
NUEVA ZELANDA	KENIA
NORUEGA	MALASIA
PORTUGAL	MÉXICO
REINO UNIDO	REP. DOMINICANA
SUECIA	SINGAPUR
SUIZA	TAILANDIA
	URUGUAY
	VENEZUELA

Elaboración propia.

La situación de crisis o no crisis de los sistemas bancarios de estos países en cada uno de los años comprendidos en el periodo temporal de estudio, da lugar a un conjunto de 458 observaciones, de las que 107 representan situaciones "malas"[3].

La explicación de estas crisis podemos encontrarlas tanto en el comportamiento de variables propias del sector financiero (aproximación micro), como en aquellas que reflejan las condiciones del entorno donde los bancos desarrollan su actividad (aproximación macro). Nuestro análisis se decanta por una aproximación micro-macro, al tomar como variables explicativas las que se reflejan en la tabla 2[4]

Tabla 2: Variables explicativas.

VARIABLE	DEFINICIÓN
PIB	Tasa de crecimiento relativo del PIB real
INFLACIÓN	Tasa de crecimiento relativo del deflactor del PIB
SALDOPRES	Saldo presupuestario sobre el PIB
IREAL	Tasa de interés de los depósitos menos inflación
IPRESIDEPOSIT	Tasa de los préstamos sobre tasa de los depósitos
M2RESERVAS	Ratio de M2 sobre reservas exteriores
TASACRED	Tasa de crecimiento del crédito real
CREDPIB	Crédito sobre el PIB
CRECDEPOSIT	Tasa de crecimiento de los depósitos reales
LIQBANK	Ratio de liquidez bancaria (reservas sobre activos)

Elaboración propia.

Las cuatro primeras son variables macroeconómicas, mientras que las seis restantes corresponden a variables más específicas del sector financiero.

Árbol de clasificación

Utilizando la función de impureza de Gini , las probabilidades a priori observadas en la muestra e igual coste de clasificación errónea para ambos grupos, obtenemos trece árboles cuyos costes asociados presentamos en la tabla 3

[3] Para la clasificación de un país con sistema bancario en crisis hemos utilizado los trabajos de Caprio y Klingebiel (2003), además, en caso de duda, se han tenido en cuenta las muestras de Demirgü**Ç**-Kunt y Detragiache (1997) y Glick y Hutchison (1999).

[4] La elección de las diez variables potencialmente explicativas de las crisis de los sistemas bancarios ha estado condicionada por la significatividad que han mostrado en trabajos anteriores y la disponibilidad de datos para los países de la muestra.

Tabla 3: Árboles construidos con el algoritmo CART.

Árbol	Número de Nodos terminales	Coste de cross validación	Coste de resustitución
1	30	1.234 +/- 0.088	0.084
4	24	1.159 +/- 0.084	0.103
5	21	1.159 +/- 0.084	0.140
6	20	1.112 +/- 0.083	0.159
7	18	1.093 +/- 0.080	0.215
8	9	1.103 +/- 0.078	0.477
9	8	1.056 +/- 0.072	0.514
10	7	1.075 +/- 0.072	0.561
11	5	1.019 +/- 0.072	0.673
12	3	0.953 +/- 0.070	0.804
13	1	1.000 +/- 0.000	1.000

Elaboración propia.

Obviamente, cuanto mayor es el número de nodos de un árbol mayor es su poder clasificatorio en la muestra original. Así, el árbol de 30 nodos terminales presenta un porcentaje de clasificación correcta en la muestra original del 98,035%, con una sensibilidad del 97,196% y una especificidad del 98,291%. Sin embargo, estos porcentajes disminuyen considerablemente cuando realizamos la validación cruzada con k=10. Pasan a ser del 71,179% para el total, con un 53,271% de sensibilidad y un 76,638% de especificidad. Ha de observarse el mayor descenso precisamente en el grupo de mayor interés.

Esto, unido al hecho de que al considerar un árbol de gran tamaño, una de las utilidades de la técnica, como es la fácil interpretación de los resultados, se pierde, como puede observarse en el anexo 1, sonde se presenta el gráfico del árbol y la importancia relativa de cada variable explicativa en la construcción del mismo, nos ha llevado a la poda del mismo siguiendo el criterio 0SERULE y a la consideración del árbol de tres nodos terminales.

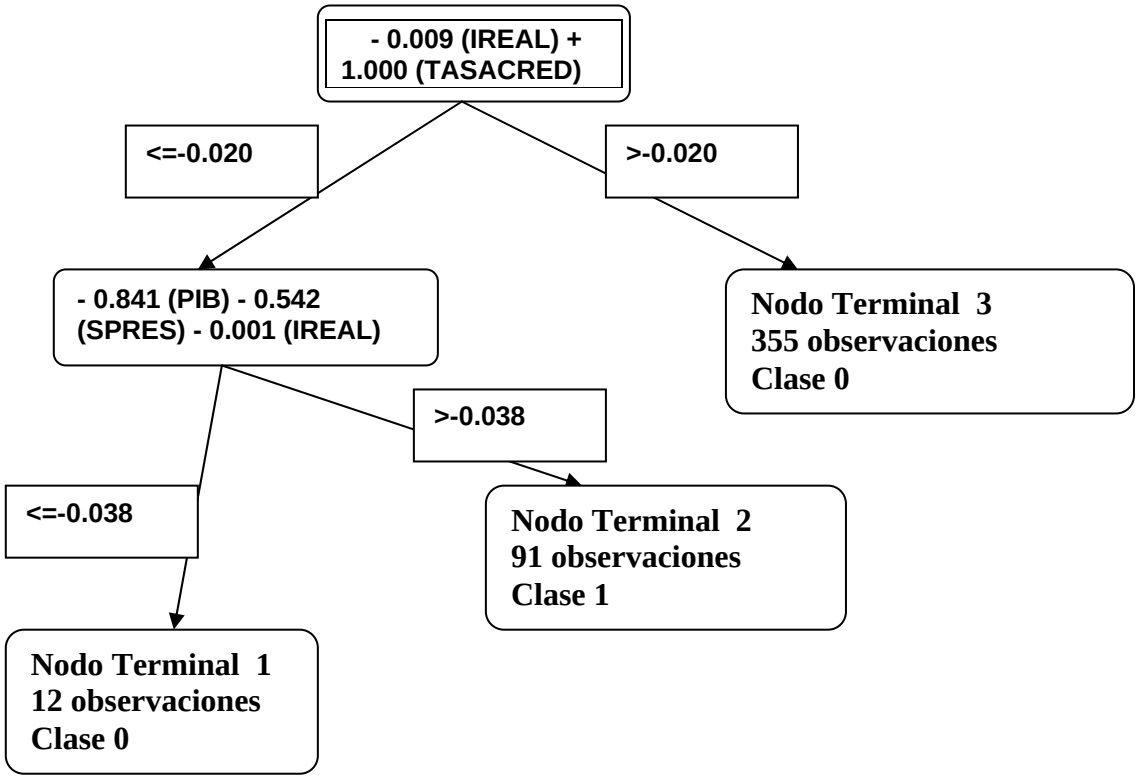
Este árbol, que se presenta en el gráfico 1, disminuye los porcentajes de clasificación correcta en la muestra original, pero los mantiene dentro de unos intervalos aceptables en la muestra de validación, ofreciendo, además, una más clara interpretación de los resultados. En la tabla 4 presentamos dichos porcentajes y la importancia relativa de cada variable en la construcción del árbol.

Tabla 4: Porcentajes de clasificación correcta del árbol seleccionado.

	Muestra original	Muestra de validación	Importancia relativa de cada variable
Porcentaje de clasificación correcta total	81,223%	77,729%	TASACRED100,00 PIB 90,14 SPRES 35,23 CDEPOSIT 32,95 IREAL 4,43 INFLACI 4,35 IPRESIDE 0,00 M2RESER 0,00 CREDPIB 0,00 LIQBANK 0,00
Especificidad	90,028%	90,313%	
Sensibilidad	52,336%	36,449%	

Elaboración propia.

Gráfico 1: Árbol de clasificación seleccionado.



Elaboración propia.

Análisis Logit

Aplicando el análisis logit a los datos de partida obtenemos los resultados que se recogen en la tabla 5

Tabla 5: Porcentajes de clasificación correcta con el análisis logit.

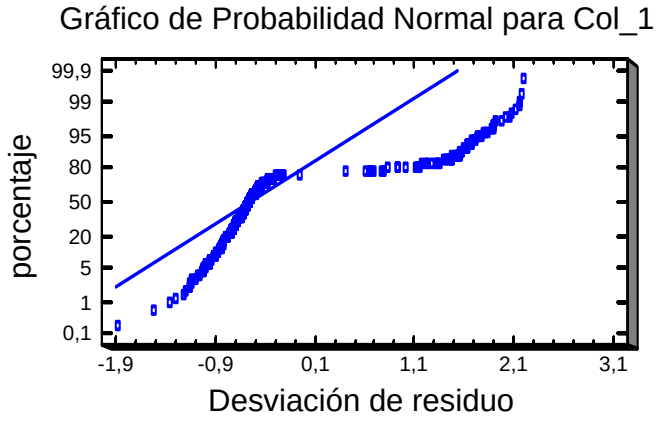
	Muestra original	Modelo	Coeficientes de las variables en la función z
Porcentaje de clasificación correcta total	69,87%	p(crisis) = exp(z)/(1+exp(z)) donde z es una combinación lineal de las variables explicativas	PIB -17,055 INFLACIÓ -4,565 SPRES -9,709 IREAL ,003 IPRESIDE -,795 M2RESERV -,002 TASACRED -2,995 CREDPIB -,238 CDEPOSIT 1,727 LIQBANK ,133 Constante ,044
Especificidad	70,655%		
Sensibilidad	67,29%		

Elaboración propia.

Con el valor del estadístico que mide la bondad del ajuste del modelo, $X^2 = 4{,}80895$, significativo al 90%.

Observamos cómo el porcentaje de clasificación correcta es menor que el obtenido con el árbol de clasificación seleccionado. Además, el modelo no verifica la hipótesis de normalidad de los residuos, como lo muestra el gráfico 2

Gráfico 2: Desviaciones de los residuos en la regresión logística.



Elaboración propia.

Con objeto de poder validar los resultados que se obtienen mediante el modelo logit, se ha repetido el análisis considerando aleatoriamente 300 de las observaciones disponibles

para construir el modelo y dejando las 158 restantes para validarlo. Los resultados, en cuanto a porcentajes de clasificación correcta se muestran en la tabla 6.

Tabla 6: Porcentajes de clasificación correcta con análisis logit validado.

	Muestra original	Muestra de validación	Modelo	Coeficientes de las variables en la función z
Porcentaje de clasificación correcta total	67,67%	62,03%	p(crisis) = exp(z)/(1+exp(z)) donde z es una combinación lineal de las variables explicativas	PIB -16,511 INFLACIÓ -3,918 SPRES -9,219 IREAL -0,019 IPRESIDE -0,527 M2RESERV 0,067 TASACRED -2,461 CREDPIB -0,097 CDEPOSIT 1,946 LIQBANK 0,176 Constante 0,404
Especificidad	68,89%	81,25%		
Sensibilidad	64,00%	57,14%		

Elaboración propia.

Con el valor del estadístico que mide la bondad del ajuste del modelo, $X^2 = 2,26$, significativo al 90%.

Vemos como el porcentaje de clasificación correcta total disminuye en la muestra de validación, volviéndose inestable respecto a la clasificación correcta de los dos grupos considerados.

El modelo identifica, además 22 residuos atípicos, lo que nos obligaría a un replanteamiento de la validez de las hipótesis adoptadas.

Con respecto a las variables observamos que el modelo identifica como significativas a PIB, INFLACIÓN, SPRES Y TASACRED, que coincide, salvo INFLACIÓN, con las variables de mayor importancia en la construcción del árbol seleccionado.

Conclusiones

Los numerosos episodios de crisis de los sistemas bancarios ocurridos en la última década del siglo pasado, han producido una ingente cantidad de datos, que han permitido a los investigadores profundizar en la búsqueda de modelos explicativos de las mismas.

Entre ellos los algoritmos de “partición recursiva binaria” ó “árboles de clasificación” se muestran como una alternativa aceptable al tradicional modelo logit de respuesta cualitativa.

Frente a estos presentan, entre otras, dos ventajas fundamentales: su fácil interpretación y la no exigencia de hipótesis “a priori” sobre las variables explicativas.

En este trabajo, donde hacemos una aplicación empírica de ambas técnicas, considerando una muestra de 40 países y diez variables potencialmente explicativas de las mismas, incluyendo variables macroeconómicas o de entorno donde se desarrolla la actividad bancaria, y ratios específicos del sector financiero, y donde conocemos la situación de crisis o no del sistema bancario de cada país para los años comprendidos entre 1988 y 2000, que es el periodo temporal de nuestro estudio, observamos como, unido a las ventajas que apuntábamos anteriormente, los porcentajes de clasificación correcta, eligiendo adecuadamente el árbol, son superiores a los obtenidos con el método logit clásico. Esta circunstancia se mantiene incluso cuando se validan los resultados de ambas técnicas con muestras distintas de la original. Se ofrece, así, una alternativa metodológica válida al análisis logit para la detección de sistemas bancarios en crisis.

Bibliografía

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

Breslow, L.A.; Aha, D.W. (1997):”Simplifying Decision Trees: A Survey” Knowledge Engineering Review, 12, pp:1-40.

Caprio, G.; Klingebiel, D. (2003): “Episodes of Systemic and Borderline Financial Crises”, Banco Mundial, mimeo.

DemirgüÇ-Kunt, A.; Detragiache, E. (1997): “The Determinants of Banking Crises: Evidence from Developing and Developed Countries”, FMI WP/97/106, Washington.

DeVaney, S. A. (1994): “The Usefulness of Financial Ratios as Predictors Of Households Insolvency: Two Perpectives” Financial Counseling and Planning, 5, pp: 5-24.

Feldesman, M.R. (2002): “Classification Trees as an alternative to linear Discriminant Analysis”American Journal Physiology Anthropology 119(3), pp:257-275.

Feldman, D.; Gross, S. (2004): “Mortgage Default: Classification Trees Analysis” Working Papers, The University of New South Wales. School of Banking and Finance and National Science Foundation.

Glick, R.; Hutchison, M. (1999): “Banking and Currency Crises: How Common are Twins?”, Center for Pacific Basin Monetary and Economic Studies, WP PB99-07, Banco de la Reserva Federal de San Francisco.

Last, M. (2004):”A Compact and Accurate Model for Classification” IEE Transactions on Knowledge and Data Engineering, vol 16, 2, pp:203-215.

Lim, T.; Loh, W.;Shih, Y.(2000):”A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Cxlassification Algorithms”. Machine Learning 40, pp: 203-228.

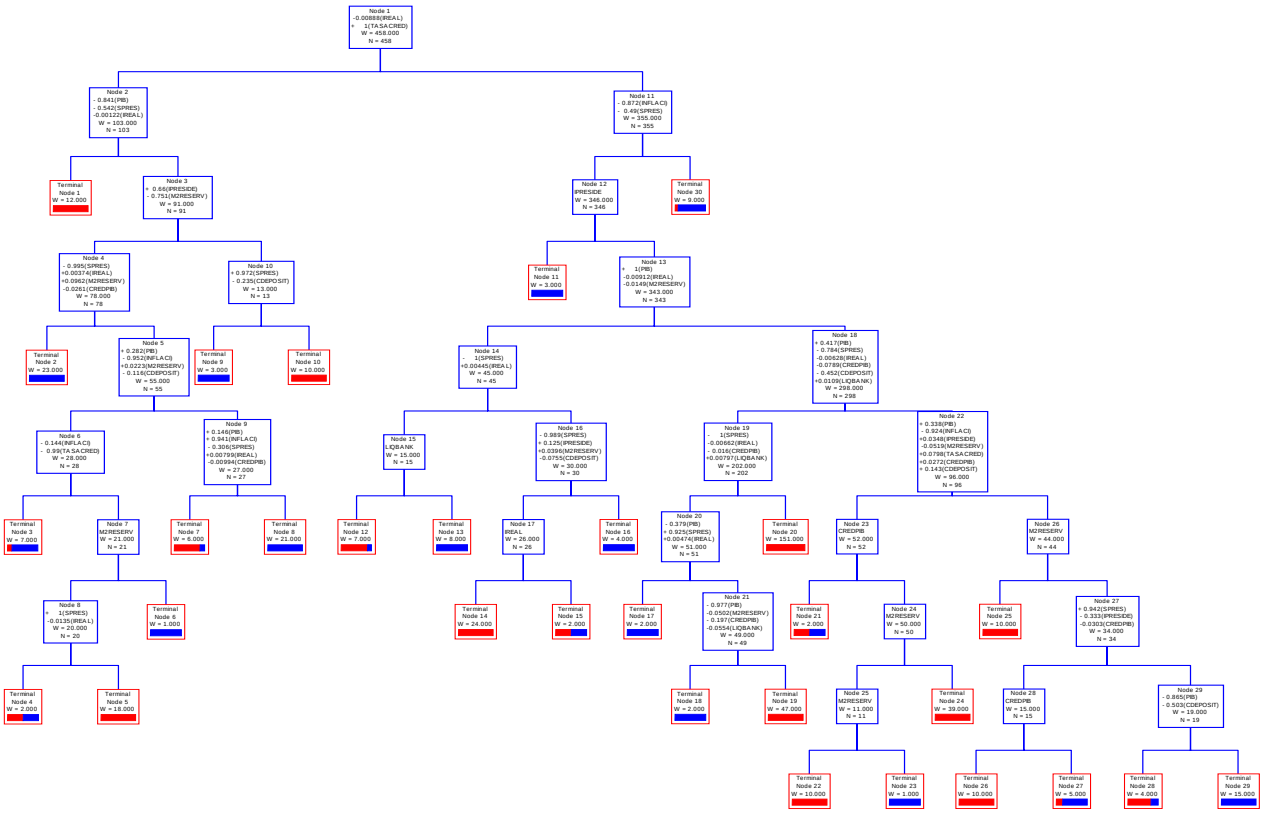
Myles, A.J.; Feudale, R.N.; Liu, Y.;Woody, N.A.; Brown, S.D. (2004):”An Introduction to Decision Tree Modeling” Journal of Chemometrics, 18, pp: 275-285.

Provost, F.; Kolluri, V.(1999):”A Survey of Methods for Scaling Up Inductive Algorithms” Data Mining and Knowledge Discovery 2, pp: 131-169.

Shih, Y. (1999): “Families of Splitting criteria for Classification Trees” Statistics and Computing vol. 9, pp: 309-315.

(2001) Statistics and Probability Letters, vol. 54, pp: 341-345.

Anexo 1. Gráfico del árbol de 30 nodos terminales e importancia relativa de las variables.



TASACRED	100,00
PIB	97,2
M2RESERV	80,69
CREDPIB	74,04
SPRES	73,77
IPRESIDE	68,16
LIQBANK	63,96
INFLACI	62,24
IREAL	61,19
CDEPOSIT	35,43

Elaboración propia