Revista Mexicana de Economía y Finanzas Nueva Época
Volumen 14 Número 3, Julio-Septiembre 2019, pp. 309-328
DOI: https://doi.org/10.21919/remef.v14i3.349

Evaluación de riesgos con Data Mining: el sistema financiero español

José Alejandro Fernández Fernández[1]
Universidad ESERP, España
Virginia Bejarano Vázquez
Universidad UNED, España
Juan Antonio Vicente Virseda
Universidad UNED, España

(Recepción: 14/octubre/2018, aceptado: 22/enero/2019)

Resumen

El objetivo de este trabajo, basado en técnicas de Data Mining, es llegar a identificar el mejor método de predicción de riesgos para el sistema bancario español, teniendo en cuenta tanto sus características específicas, como la situación económica de España en el período objeto de estudio. Para ello, se definen, en primer lugar, catorce ratios que permiten identificar, en términos de riesgos, la situación de los bancos y cajas de ahorros españoles durante el período examinado. Mediante una técnica de reducción de dimensiones con la que se simplifica la interpretación de resultados, se obtienen cuatro factores latentes sobre los que se evalúa, junto con cuatro variables macroeconómicas adicionales,un conjunto de algoritmos de Data Mining, siendo seleccionado finalmente el arbol CHAID, a diferencia de trabajos previos, en los que nunca se había llegado a proponer la aplicación de técnicas de Data Mining y Machine Learning en la identificación de situaciones de riesgo en la industria bancaria española. Una limitación de este trabajo ha sido la imposibilidad de incorporar variables regulatorias, por tratarse de información reservada que, de haber estado disponible, nos habría permitido incorporar una nueva dimensión en la predicción de riesgos.

Clasificación JEL: G21, G33, M41
Palabras clave: Data Mining, Machine Learning, métodos de clasificación, predicción de riesgos, solvencia

Risk assessment with Data Mining: the Spanish financial system

Abstract

The objective of this work, based on Data Mining techniques, is to identify the best risk prediction method for the Spanish banking system, taking into account its specific characteristics and the economic situation of Spain during the period under study. For this purpose, first of all, fourteen ratios are defined in order to identify, in terms of risks, the situation of Spanish banks and savings banks during the period under review. Through a technique of reduction of dimensions which simplifies the interpretation of results, four latent factors are obtained on which are evaluated, together with four additional macroeconomic variables, a set of algorithms of Data Mining, being finally selected the CHAID tree, unlike previous works, in which it had never come to propose the application of techniques of Data Mining and Machine Learning in the identification of situations of risk in the Spanish banking industry. One limitation of this work has been the impossibility of incorporating regulatory variables, due that this information is confidential,

[1]Correo electrónico: alejandr0fernandez@hotmail.com Dirección: Calle Fernando el Católico, 3 6B izquierda. Madrid (España). Teléfono:0034 606907537
*Sin fuente de financiamiento declarada para el desarrollo de la investigación

Abstract

otherwise, we would have been able to incorporate a new dimension in the prediction of risks.
JEL Classification: G21, G33, M41
Keywords: Data Mining, Machine Learning, classification methods, risk prediction, solvency

1. Introducción

A partir de 2008, coincidiendo con el inicio de la última crisis financiera, se aprobaron diversas normativas en distintos ámbitos -nacional e internacional- con el objetivo común de permitir actuar anticipadamente sobre aquellas entidades que presentaran problemas de solvencia. Concretamente, en el ámbito de la Unión Europea se debe destacar la Directiva 2014/59/UE, de 15 de mayo de 2014, en la que se establecen directrices generales en materia de reestructuración y resolución bancaria. En España, la Ley 11/2015, de 18 de junio, de recuperación y resolución de entidades de crédito y empresas de servicios de inversión (Ley 11/2015), supuso la transposición de dicha directiva, siendo uno de sus principales propósitos evitar la toma de estrategias excesivamente arriesgadas, o demoras en la ejecución de recapitalizaciones necesarias y, en general, tratar de limitar la adopción de políticas de riesgos que pudieran acentuar problemas existentes.

Este estudio se inicia con una revisión de los trabajos previos realizados en la materia aplicando metodologías similares o, directamente, modelos estadísticos con nuestro mismo objetivo; la identificación de entidades con problemas de solvencia, incluidos aquellos trabajos que toman únicamente variables microprudenciales, si bien en este trabajo se incluyen además variables macroeconómicas, representativas de la situación económica española en el período objeto de análisis.

La investigación propiamente dicha se inicia con un análisis de los datos obtenidos de las cuentas anuales consolidadas de todos los bancos y cajas de ahorros que integran el sistema financiero español para el período comprendido entre los años 2005 y 2012, ambos inclusive, previa definición y cálculo de catorce ratios, a los que se ha aplicado un análisis factorial que nos ha permitido obtener cuatro factores subyacentes, capaces de explicar la evolución de la situación económico-financiera de cada entidad a lo largo del período analizado.

El estudio se completa incorporando los cuatro factores obtenidos, junto con una selección de cinco variables macroeconómicas, a los siguientes modelos predictivos: árboles de decisión (CRT, QUEST, y CHAID) redes neuronales (perceptrón multicapa y de funciones de base radial) y regresión logística y nominal, concluyendo que todos los modelos podrían resultar equivalentes en términos de predicción de riesgos (detección de entidades con problemas), si se toma como criterio el área bajo la curva ROC. Sin embargo, se ha optado por seleccionar el modelo CHAID (en su versión exhaustiva) por ser el algoritmo que obtiene la mayor precisión y por su sencillez, incorporando únicamente tres variables predictoras; dos microeconómicas y una macroeconómica.

2. Revisión de la Literatura

Desde los trabajos pioneros en materia de predicción de quiebra de Sinkey (1975) o Altman et al. (1977), que aplican un análisis múltiple discriminante (MDA), puede encontrarse en la literatura numerosas investigaciones, que van desde la más sencilla, como la realizada por Yiqiang et. al. (2011) en Estados Unidos, basándose en el análisis estadístico univariante de medias para distinguir entre bancos quebrados y no quebrados durante la crisis de 2007, -también utilizado por Cox y Wang (2014) junto con un análisis discriminante-, hasta modelos más complejos, como el desarrollado por De Andrés et al. (2011) para

la previsión de quiebra en empresas españolas, combinando el agrupamiento difuso con la Multivariate Adaptive Regression Splines (MARS), obteniendo un rendimiento superior en la predicción de la quiebra bancaria en España, al de las redes neuronales o al del análisis discriminante aplicado por Serrano y Martín del Brio (1993) en un trabajo previo. Todos los trabajos anteriores adolecen de una limitación importante; no definir las características latentes del sector sobre el que se realiza el análisis. Por ello, nuestra investigación se inicia con un análisis factorial que nos ha permitido obtener dichas características o dimensiones latentes.

De toda la literatura examinada en materia de análisis factorial destaca el trabajo de Cambas et al. (2005) sobre predicción de quiebras bancarias en Turquía, en el que concluyen que dicha técnica puede utilizarse como herramienta complementaria a la metodología CAMEL. Concretamente, a partir de tres factores (adecuación de capital, estructura de ingresos-gastos y liquidez), dichos autores obtienen las características financieras latentes del sistema bancario turco para, a continuación, predecir situaciones de quiebra aplicando el modelo discriminante logit y probit. En nuestro trabajo, el análisis factorial, aplicado a catorce variables microeconómicas predefinidas, nos permite obtener las características financieras principales del sistema bancario español: solvencia, liquidez, rentabilidad y tamaño, siendo éste uno de los elementos que dota de originalidad a nuestra investigación, junto con la realización de un análisis Benchmark con el que se demuestra la superioridad de los resultados obtenidos con la aplicación de dicho análisis, frente a los que se habría alcanzado de no haberlo aplicado.

Otro de los elementos característicos de nuestro trabajo reside en la incorporación de variables macroeconómicas en la fase de análisis y selección del mejor método de predicción de riesgos bancarios. Entre los trabajos que incorporan variables "macro" destaca el de Betz et al. (2013), sobre predicción de quiebra bancaria en distintos países de la UE, en el que incluyen la inflación, la variación del PIB, el flujo de crédito del sector privado, la Deuda Pública y la posición de inversión internacional, dentro de un modelo logit. En nuestro trabajo también se han incorporado variables macroeconómicas, por estimar que con ello se enriquece el análisis, al reflejar tanto la situación económica del área geográfica objeto de estudio, en nuestro caso España, como la existencia y posible influencia de desequilibrios macroeconómicos.

En cuanto a los métodos de predicción de quiebra, son numerosos los trabajos realizados en todas las industrias, basándose en la comparación de diversos modelos de predicción y búsqueda de aquél con mayor capacidad predictiva. Así, en el caso de la industria bancaria, objeto de este trabajo, destaca el de Boyacioglu et al. (2009) que, aplicando las redes neuronales, las máquinas de vectores de soporte y métodos estadísticos multivariantes (análisis discriminante multivariante, análisis de conglomerados k-medias y análisis de regresión logística) para predecir la quiebra bancaria en Turquía, llegan a la conclusión de que las redes neuronales perceptrón multicapa son superiores al resto de modelos analizados. En línea con el trabajo anterior, Le y Viviani (2017), que centran su investigación en los Estados Unidos, combinan técnicas estadísticas tradicionales (análisis discriminante y regresión logística) con técnicas de aprendizaje automático (Red neuronal artificial, máquinas de vector soporte y los k-vecinos más próximos) para concluir que la red neuronal artificial y el método de los k-vecinos más próximos son los más precisos. Se trata de un estudio similar al presentado en este trabajo, en la medida en que, a partir de los datos obtenidos de los informes financieros, calculan 31 ratios con los que obtienen como variables microeconómicas más relevantes la eficiencia operativa, la rentabilidad y la liquidez, frente a los resultados obtenidos en el nuestro, en el que la característica más relevante del sistema financiero español, a efectos de predicción de problemas bancarios, es la solvencia.

Entre los trabajos que obtienen los mejores resultados con las redes neuronales, cabe citar, además del de Le y Viviani (2017), el de Boyacioglu et al. (2009) y el de Madireddi

y Vadlamani (2011). En todos estos trabajos, tras realizar una revisión de los métodos aplicados a la bancarrota, se concluye que las redes neuronales y las máquinas de vector soporte son los métodos más precisos. Sin embargo, otros trabajos, como el de Alaka et al. (2017), advierten que los resultados obtenidos con modelos de redes neuronales y máquinas de vector soporte asignan a las variables ponderaciones/coeficientes ilógicos y difíciles de interpretar. Para tratar de salvar dicha limitación, en nuestro trabajo se ha optado por testar todos los modelos posibles, incluidos los árboles de decisión.

Entre los trabajos que avalan los buenos rendimientos de los árboles de decisión, aunque no se circunscriben a la industria bancaria, cabe citar el realizado por Olson et al. (2012), en el que comparan distintos métodos de predicción de quiebras en compañías norteamericanas, en el período 2005 a 2009, concluyendo que los árboles de decisión son más precisos que las redes neuronales y las máquinas de vectores de soporte, si bien, con el inconveniente de que producen más nodos de los deseables. Por ello, justifican un ajuste en los mismos, a fin de generar reglas más manejables. Igualmente, Aktan (2011) examina distintos métodos de predicción de quiebra para empresas industriales en Turquía (Naive Bayes, Red Bayesiana, k-vecinos, redes neuronales, máquinas de vector soporte y árboles C4.5, CHAID y CRT), y concluye que el árbol de decisión CRT permite obtener los mejores resultados. Chen (2011), por su parte, centrándose en compañías cotizadas en la bolsa de Taiwán para el período 2000 a 2007, compara los árboles de decisión con la regresión logística y concluye que los primeros presentan un mejor rendimiento como método de predicción de quiebra que la regresión logística.

Brezigar-Masten y Masten (2012) realizan un estudio sobre predicción de quiebra en las compañías eslovenas, utilizando los árboles de decisión y el modelo Logit, destacando como ventaja de los árboles de decisión la capacidad para generar reglas de decisión fácilmente comprensibles, característica que no es compartida por muchos métodos de inteligencia artificial. A la lista anterior se podrían sumar los trabajos de Koyuncugil y Ozgulbas (2009), Serhan y Ozgulbas (2012), Gepp y Kumar (2015), Liviu et al. (2015) o Irimia-Dieguez et. al. (2015).

Dadas las características del sistema bancario español, en este trabajo se llega a la conclusión de que a través de una metodología sencilla y fácil de explicar, como es un árbol de decisión, se puede proporcionar un sistema que permita al supervisor identificar aquellas entidades que podrían encontrarse próximas a la quiebra o con serios problemas de solvencia.

3. Metodología

Para identificar el mejor método de predicción de situaciones de riesgo en las entidades de depósito del sistema bancario español se ha aplicado, en una primera etapa, el análisis factorial, metodología que nos ha permitido identificar las características latentes (variables microeconómicas) de los bancos y cajas de ahorro españoles. A continuación, se han aplicado los siguientes modelos de predicción de riesgos: red neuronal preceptrón multicapa y funciones de base radial, árboles de decición CRT, QUEST, y CHAID (versión eshaustiva), regresión logística y nominal, a las cuatro características latentes obtenidas y a otras cuatro variables macroeconómicas, a fin de identificar aquel modelo que muestre mayor capacidad predictiva en términos de riesgos bancarios.

3.1 Análisis factorial

En cualquier estudio que pretenda analizar la situación económico-financiera de una entidad al objeto de identificar posibles problemas o, simplemente, desequilibrios en alguna de las dimensiones indicadas, se deben identificar cuatro magnitudes fundamentales: solven-

cia[2], liquidez[3], rentabilidad[4] y tamaño. Con objeto de medir cada uno de estos conceptos, se ha seleccionado una batería de indicadores (ratios) representativos de cada uno de ellos, teniendo en cuenta la doctrina existente en la materia. A partir de dichas variables, se han obtenido cuatro factores, también denominados "dimensiones latentes", que recogen las características de los indicadores utilizados para su cálculo, obteniendo con ello una interpretación más sencilla de la realidad objeto de estudio.

3.1.1 Configuración experimental

En este trabajo, a diferencia de otros realizados sobre una muestra, se ha partido del censo de todos los Bancos y Cajas de Ahorro que integran el sistema financiero español para el período comprendido entre los años 2005 Y 2012 (ambos inclusive). Todos los datos se han obtenido directamente de las cuentas anuales consolidadas publicadas por la Asociación Española de Banca (AEB) y por la Confederación Española de las Cajas de Ahorro (CECA), disponibles en sus respectivas páginas web[5]. Con esos datos se han calculado catorce ratios previamente definidos como se muestra en la tabla 1. El análisis factorial se ha realizado a partir de los resultados obtenidos, para cada una de las entidades del censo, de los catorce ratios definidos teniendo en cuenta los objetivos del trabajo, consiguiendo su agrupación en los cuatro factores siguientes: solvencia, liquidez, tamaño y rentabilidad, que han permitido identificar las características financieras latentes del sistema bancario español durante el período objeto de análisis.

Se definen y justifican, a continuación, los valores medios obtenidos para cada uno de los factores latentes obtenidos. Concretamente, en los gráficos 1 y 2, se muestra la evolución, a lo largo de los ocho años objeto de estudio, de los cuatro factores identificados, distinguiendo por bancos (gráfico 1) y cajas de ahorro (gráfico 2). Lo más significativo en el caso de los bancos es el comportamiento de los factores rentabilidad y liquidez, que muestran un acusado descenso a la baja. Así, por ejemplo, coincidiendo con los valores medios del ratio (10) que, de acuerdo con los resultados de la matriz de componentes rotados (ver tabla 4) es el que contribuye en mayor medida a explicar la rentabilidad de los bancos, a un tamaño medio de los bancos similar a lo largo de la serie temporal, el resultado de explotación alcanza un pico de crecimiento en el año 2006, para experimentar un acusado descenso en 2008 que se mantiene en 2009 y seguir reduciéndose el resto de años. El factor solvencia, por su parte, mantiene un comportamiento más o menos homogéneo a lo largo de la serie. El tamaño es el único factor que, tras una ligera caída hasta el año 2007, ha experimentado leve crecimiento a lo largo del resto de la serie temporal.

En el caso de las cajas de ahorros el comportamiento de todos los factores es similar, en la medida que todos experimentan una suave caída, en cualquier caso, menos acusada que en los bancos, a excepción de la rentabilidad, que sufre una brusca caída en 2012, circunstancia que se podría explicar en gran medida, de nuevo, por el comportamiento del ratio número (10), que en dicho año pasa de un valor medio de "0,0" a un valor negativo.

En el gráfico 3, por su parte, se muestra el diferencial de bancos vs. cajas de ahorros. Como puede observarse, los bancos muestran, a lo largo de todo el período, una mayor solvencia, liquidez y tamaño, pero son menos rentables. En 2012, la rentabilidad de los bancos se sitúa también por encima, debido a la fuerte caída experimentada por las cajas de ahorro.

[2]Entendida como la "capacidad de una empresa para hacer frente a la devolución de las obligaciones financieras asumidas" (Corona et. Al. (2017: 116).

[3]Indicador de las posibilidades para generar los recursos líquidos necesarios para atender los compromisos asumidos en el desarrollo de la actividad típica de la empresa (Corona et. Al. (2017: 116).

[4]Capacidad de una empresa para generar ganancias o rendimientos.

[5]https://www.aebanca.es/
http://www.ceca.es/

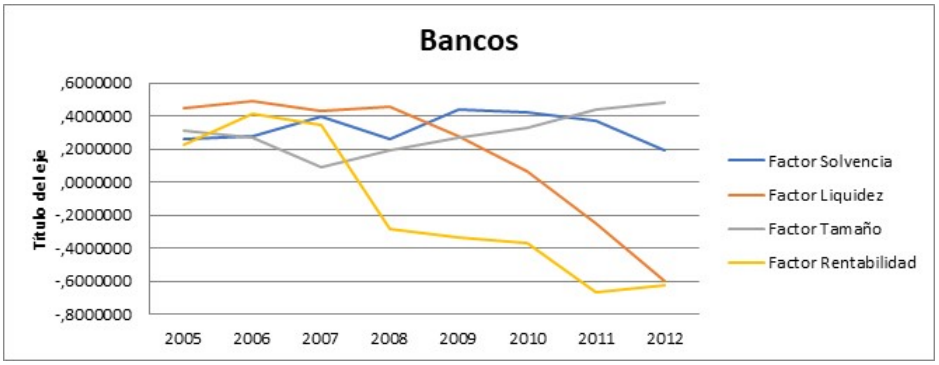

Gráfico 1. Evolución factores en bancos
Fuente: Elaboración propia.

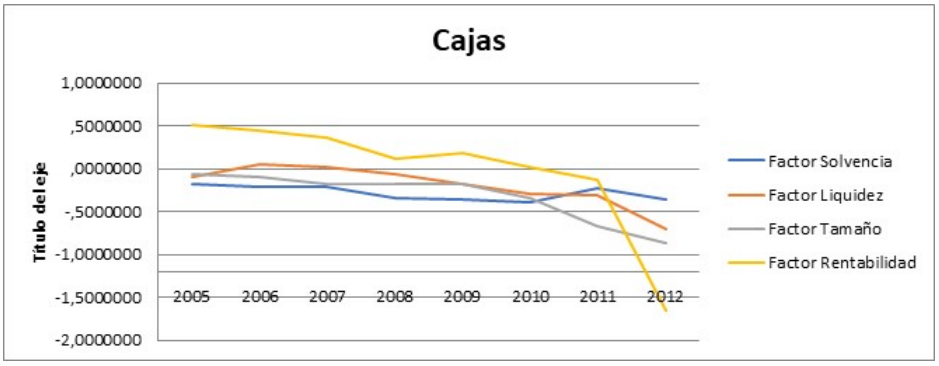

Gráfico 2. Evolución factores en cajas de ahorros
Fuente: Elaboración propia.

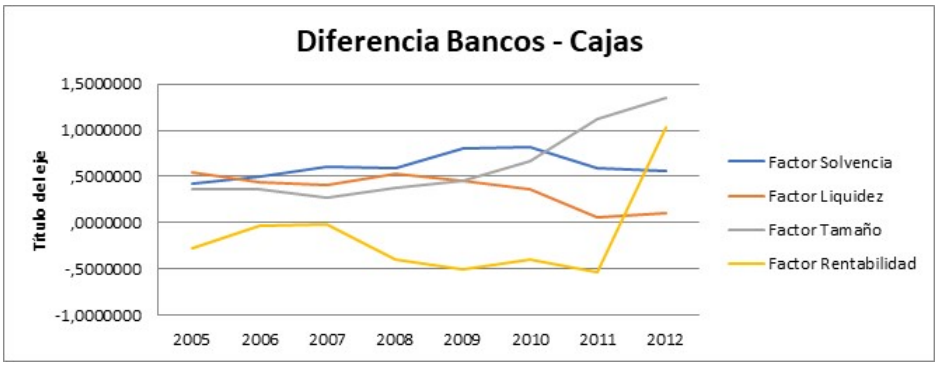

Gráfico 3. Evolución factores bancos vs. cajas de ahorros
Fuente: Elaboración propia.

En el Anexo de este trabajo se muestra la estadística descriptiva de las variables utilizadas en el análisis factorial. Concretamente, en la tabla 1 del indicado Anexo se calcula, para cada uno de los catorce ratios, el valor medio y la desviación típica en cada uno de los años objeto de estudio (2005 a 2012) de las cajas de ahorro españolas. En la tabla 2 se calculan los mismos estadísticos para el caso de los bancos.

Tabla 1. Análisis factorial. Variables utilizadas

Microeconómicas	
Variable	**Descripción**
(1) Patrimonio neto Activo	Mide la capacidad de una entidad para soportar pérdidas potenciales o disminuciones inesperadas en el valor de sus activos, sin que los acreedores o depositantes sufran pérdidas.
(2) Patrimonio neto Pasivo	Medida de solvencia y apalancamiento de una entidad. Cuanto mayores sean los valores que toma el ratio mejor será la solvencia de la entidad y, al tiempo, menor será su apalancamiento.
(3) Patrimonio neto Pasivo + cuentas orden	Medida de solvencia y apalancamiento, en la que, además, se incorpora aquellas cuentas representativas de obligaciones potenciales de la entidad, como garantías financieras o compromisos de compra, entre otros (cuentas de orden).
(4) Pasivos financieros a coste amortizado Activo	Medida de solvencia de una entidad que indica la porción de la estructura financiera ajena de una entidad que no tiene carácter especulativo (pasivos financieros a coste amortizado como los depósitos, los pasivos subordinados y los débitos por valores negociables). Cuanto mayor sea el valor de este ratio mayores serán las obligaciones de la entidad, con la consiguiente erosión de su solvencia.
(5) Inversión crediticia Activo	Medida de liquidez de una entidad. Se asume que cuanto mayor sea este ratio menor será la liquidez, al estar invertidos los recursos financieros en préstamos y valores a largo plazo.
(6) Crédito a clientes Activo	Mide la proporción de la inversión total de la entidad que está materializada en "crédito a la clientela" independientemente de la cartera en la que esté clasificado dicho crédito: i) negociación (presenta principalmente riesgo de mercado); ii) otros activos financieros a valor razonable con cambios en pérdidas en ganancias; e iii) inversiones crediticias. Permite intuir la política de gestión del riesgo de crédito de una entidad observando la composición y ponderación de estas carteras, principalmente la de inversión crediticia, por ser la de mayor tamaño y representar el negocio tradicional de las entidades de crédito. De manera que si este ratio toma valores próximos a los de la variable (5) la actividad de la entidad estará centrada en el negocio tradicional bancario, al tener un peso residual las carteras de "negociación" y de "otros activos financieros a valor razonable con cambios en pérdidas en ganancias".
(7) Depósitos de clientes Préstamos concedidos	Este ratio permite distinguir la financiación con depósitos de clientela, más estable al estar protegida por el Fondo de Garantía de Depósitos (FGD) y no quedar expuestos al riesgo de crédito, de la financiación mayorista que presenta el riesgo de no ser renovada y que durante la crisis fue la causa de numerosos problemas de liquidez en numerosas entidades, que recurrieron a ella para financiar su activo.
(8) Activos financieros disponibles para la venta Activo	Medida de liquidez de una entidad, en la medida que la cartera de activos disponibles para la venta recoge instrumentos de deuda y de capital que suelen negociarse en mercados líquidos.
(9) Resultado del ejercicio Activo	Mide la capacidad global de una entidad para generar rendimientos con sus recursos económicos. Se trata de una medida de la calidad de la gestión de una entidad que puede identificarse con la rentabilidad económica (ROA).

(10) Rdo de explotación Activo	Mide la rentabilidad generada por el negocio recurrente. Se obtiene agregando al margen bruto los gastos de administración, de personal, otros gastos generales de administración, amortización, dotación de provisiones (neto) y las pérdidas por deterioro de activos financieros (neto).
(11) Margen bruto Activo	Medida de rentabilidad más próxima al núcleo de la actividad bancaria sobre la que se asume que, en principio, los directivos tendrán una menor liberalidad para operar, al no incluir partidas como amortizaciones, gastos de administración o las pérdidas por deterioro de activos financieros.
(12) Logaritmo del activo	Partiendo de que el tamaño constituye una variable relevante, al determinar economías de escala y posibilitar la aplicación de políticas de diversificación de riesgos más activas, para medirlo, se ha tomado el logaritmo de los activos, realizando un cambio de escala. Todo ello, sin perjuicio de que un mayor tamaño incrementa el riesgo moral ("demasiado grande para caer"), tal y como se manifiesta en que, en Basilea III, el tamaño se incluye en la "exposición total" a efectos del coeficiente de solvencia, para identificar bancos con riesgo sistémico e imponerles mayores requerimientos de capital.
(13) Logaritmo de los ingresos por intereses	Medida que atiende al volumen de negocio, si bien, centrándose exclusivamente en los ingresos por intereses que, como se conoce, se encuadran dentro del negocio tradicional bancario.
(14) Costes operativos Ingresos operativos	Medida de ineficiencia, dado que cuanto mayores sean los valores que toma este ratio más ineficiente será la entidad.
Macroeconómicas[6]	
Saldo Balanza corriente	Mide la posición neta de ahorro de una economía en relación con el resto del mundo. Es el concepto económico más utilizado al analizar desequilibrios globales.
Financiación total a hogares	Variable que mide el grado de endeudamiento de las economías domésticas.
Inflación	Se trata de una variable que determina el grado de estabilidad de una economía.
Variación del PIB	Variable que permite identificar si una economía se encuentra en fase de expansión o de recesión.
Variación de la Deuda Pública	Variable medida como porcentaje del PIB, que permite analizar la evolución de la financiación del sector público en una economía.

Fuente: elaboración propia.

En cuanto a las variables macroeconómicas seleccionadas: i) saldo de la balanza por cuenta corriente; ii) financiación total a los hogares; iii) inflación; iv) variación del PIB y; v) variación de la Deuda Pública, su elección se ha basado en que permiten identificar desequilibrios internos y externos (deuda pública, saldo de la balanza de pagos o financiación a los hogares), así como variaciones coyunturales de una economía (inflación y variación del PIB). Al igual que en el caso de las variables microeconómicas, los datos corresponden a los años 2005 a 2012, ambos inclusive, y se han obtenido directamente del Boletín estadístico del Banco de España, disponible en su página web[7].

3.1.2 Ejecución del experimento

El Análisis Factorial busca obtener factores que expliquen la mayor parte de la varianza común de las variables analizadas; en nuestro caso, los catorce ratios definidos ad hoc y calculados. Se trata, por tanto, de calcular nuevas "variables ficticias" que, aunque no observables, son combinación lineal de las originales y recogen la mayor parte de la información correspondiente a las primeras. Como se puede observar en la tabla 2, el KMO indica una aceptable adecuación de los datos al modelo factorial, al tener un valor de 0,705. Por su parte, el test de esfericidad de Barlet indica una elevada correlación entre las variables originales, lo que permite concluir que es posible la reducción de la dimensionalidad.

Tabla 2. Adecuación del modelo.
KMO y test de esfericidad de Barlet.

Kaiser-Meyer-Olkin Medida de adecuación muestral		0,705
Test de Esfericidad de Bartlet	Aprox. Chi-cuadrado	9.315,945
	Df	91
	Sig.	0,000

Fuente: elaboración propia

En la tabla 3 se muestra la varianza explicada por cada uno de los factores obtenidos y su porcentaje. Cabe destacar que los cuatro primeros factores obtienen valores propios mayores que uno, consiguiendo explicar con ellos el 81,0 % de la varianza. El quinto factor es despreciado por no alcanzar un valor propio mayor que uno (0,734) indicativo de su escaso poder explicativo.

Tabla 3. Varianzas explicadas

Factor	Autovalores iniciales			Suma de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% Varianza	% Acum.	Total	% Varianza	% Acum.	Total	% Varianza	% Acum.
1	4,933	35,238	35,238	4,933	35,238	35,238	4,030	28,784	28,784
2	3,128	22,342	57,579	3,128	22,342	57,579	2,984	21,318	50,101
3	2,008	14,342	71,921	2,008	14,342	71,921	2,287	16,338	66,439
4	1,265	9,034	80,955	1,265	9,034	80,955	2,032	14,516	80,955
5	0,734	5,246	86,201						

Fuente: elaboración propia.

Para simplificar la estructura factorial se ha realizado una rotación de los ejes factoriales. De entre los diferentes procedimientos existentes, se ha utilizado el método Varimax (Kaiser, 1958), cuya finalidad es simplificar la estructura factorial maximizando la varianza de los coeficientes factoriales al cuadrado para cada factor. En la Tabla 4 se recoge la matriz de componentes rotadas, que representa la estructura factorial resultante de la rotación indicada. En ella se puede observar cada una de las dimensiones en que quedan definidas las variables originales (ratios), así como la capacidad de cada una de estas variables a la hora de explicar el comportamiento del factor en el que ha quedado incluida.

[7] https://www.bde.es/

Tabla 4. Matriz de componentes rotados

	Componente			
	Solvencia (1)	**Liquidez (2)**	**Tamaño (3)**	**Rentabilidad (4)**
(2)	,923			
(1)	,906			
(3)	,896			
(4)	-,792			
(11)	,688			
(5)		,917		
(6)		,905		
(7)		-,818		
(8)		-,754		
(12)			-,937	
(13)			-,919	
(10)				,894
(9)				,871
(14)				-,578

Fuente: Elaboración propia.

3.1.3 Análisis y discusión de los resultados

Se analiza y discute a continuación, los resultados obtenidos para cada uno de los cuatro factores seleccionados:

a) El **primer factor**, representativo de la "SOLVENCIA" de una entidad y constituido por los ratios (1), (2), (3), (4) y (11) (ver tabla 4), ha permitido confirmar que una mayor solvencia está relacionada positivamente con un mayor peso de la financiación propia sobre la ajena, así como con un menor peso de la cartera de pasivos financieros a coste amortizado, cartera que, como se conoce, juega en contra de dicha solvencia. En sentido contrario, juegan las obligaciones posibles de una entidad ("cuentas de orden"), que actúan restando solvencia a la entidad.

   Adicionalmente, se puede afirmar que las entidades más solventes son las que presentan un mayor margen bruto sobre los activos totales medios porque, por un lado, y teniendo en cuenta que el componente principal de dicho margen es el "margen por intereses", elemento nuclear del negocio bancario tradicional, dichas entidades tienen menores costes de financiación, circunstancia que se puede justificar, por una parte, en un mayor valor del ratio (2) que, en última instancia, redunda en un menor riesgo percibido por los mercados, con el consiguiente efecto de rebaja en los costes de financiación. Adicionalmente, las entidades más solventes obtienen un mejor resultado por operaciones financieras, al presentar menores dificultades para obtener liquidez por la vía de la enajenación de carteras, circunstancia que evita tener que realizar ventas a bajos precios que puedan perjudicar dichos resultados.

b) El **segundo factor**, (ver tabla 4) representativo de la "**LIQUIDEZ**" de las entidades, confirma que cuanto mayor es la inversión crediticia de las entidades, menor es su liquidez. En sentido contrario, cuanto mayor es la proporción de los depósitos respecto a los préstamos de la clientela mayor es la liquidez de la entidad, dado que dichos depósitos constituyen una financiación más estable derivada de la garantía prestada por el Fondo de Garantía de Depósitos (FGD)[8]; téngase en cuenta que uno de los

[8]Creado por Real Decreto-ley 16/2011, de 14 de octubre, con personalidad jurídica propia y plena capacidad para el desarrollo de sus fines en régimen de derecho privado, tiene por objeto garantizar los depósitos en dinero y en valores u otros instrumentos financieros constituidos en las entidades de crédito, con los límites establecidos legalmente. Para el cumplimiento de su función de garantía de depósitos y en

problemas de las cajas de ahorro durante la crisis fue la "evaporación" de las fuentes de financiación a corto plazo obtenidas en los mercados mayoristas, confirmando el riesgo de no renovación que incorporan, presente en mercados con turbulencias. Por otra parte, cuando los préstamos de la clientela como recursos comprometidos a largo plazo son menores, mayor es el volumen de activos más líquidos.

c) El **tercer factor**, representativo de la dimensión "**TAMAÑO**" (ver tabla 4), está relacionado con el riesgo sistémico de una entidad y, por tanto, con el apoyo implícito que puede proporcionarle el Estado, sin olvidar los mayores requerimientos de capital de Basilea III.

d) El **cuarto factor** representa la dimensión "**RENTABILIDAD**" de las entidades (ver tabla 4). Por su propia definición, aquellas que presentan mayores valores en los ratios representativos de este factor, son más rentables. La única excepción es el ratio (11), que no satura con el resto "rentabilidades", posiblemente porque para el cálculo de esas rentabilidades, a diferencia del margen bruto, se toman dos magnitudes de resultados que incluyen medidas más "discrecionales" para la dirección.

3.2 Modelos de predicción

La segunda fase del trabajo se centra en la realización de un proceso de Data Mining, proceso interactivo que combina la experiencia sobre un problema dado con variedad de técnicas tradicionales de análisis de datos y tecnología avanzada de aprendizaje automático, con el objetivo de descubrir patrones y relaciones en los datos para la realización de predicciones válidas.

La minería de datos es una parte de un proceso más general que se denomina Descubrimiento de conocimiento en las bases de datos (Knowledge Discovery in Databases o KDD), si bien, en la mayor parte de la bibliografía sobre el tema el concepto de Data Mining toma el significado global del proceso.

Una definición más general de la minería de datos es referirse a ella como la extracción no trivial de la información implícita, previamente desconocida y potencialmente útil, a partir de los datos.

La determinación de estos modelos se ha realizado teniendo en cuenta las características de la muestra, en este caso un censo, y en concreto, los modelos finalmente considerados de cara a ser testados son los siguientes:

- Estadística clásica: regresión logística o modelo logit y regresión nominal.
- Redes neuronales artificiales: perceptrón multicapa y funciones de base radial.
- Árboles de decisión: CRT, QUEST Y CHAID, en su versión exhaustiva.

3.2.1 Configuración experimental

Para alcanzar el objetivo indicado más arriba, el suceso seleccionado en nuestro análisis fue que la entidad se encontrara en alguna de las situaciones siguientes:

- Participación en procedimientos de recapitalización, bien por la vía de inyecciones de capital o por la de las fusiones y adquisiciones, o
- Participación en sistemas institucionales de protección como la concesión de ayudas públicas o, directamente, mediante la intervención de los poderes públicos.

defensa de los depositantes cuyos fondos están garantizados y del propio Fondo de Garantía de Depósitos de Entidades de Crédito, el Fondo podrá adoptar medidas de apoyo a la resolución de una entidad de crédito con cargo al compartimento de garantía de depósitos. Excepcionalmente, siempre y cuando no se haya iniciado un proceso de resolución, el Fondo podrá utilizar sus recursos para impedir la liquidación de una entidad de crédito en los supuestos legalmente previstos (ver: https://www.fgd.es/es/index.html).

De acuerdo con lo anterior, el objetivo es la identificación de aquellas entidades que presenten el suceso descrito, en cuyo caso, serán incluidas en la categoría "si", correspondiente a entidades con riesgos bancarios. A la inversa, en la categoría "no" se ha incluido a todas aquellas entidades en la que el suceso no se ha presentado, es decir, entidades que ni han quebrado ni están cerca de hacerlo y que, por tanto, no presentan riesgos bancarios.

3.2.2 Ejecución del experimento

En la tabla 5 se muestran los resultados obtenidos con los métodos de predicción utilizados. Concretamente, la precisión o porcentaje de acierto global en cada una de las categorías de la variable respuesta, así como el área bajo la curva ROC, curva que representa la sensibilidad en el eje de ordenadas y 1-especificidad en el eje de abscisas y que se puede interpretar como un gráfico de las tasas de verdaderos positivos frente a las de falsos positivos.

Tabla 5. Clasificación de los modelos

Técnica	Clase Si (%)	Clase No (%)	Total (%)	Área ROC
Árbol CHAID exhaustivo	73,5 %	94,9 %	92,2 %	0,899
Árbol CRT	60,3 %	96,4 %	91,8 %	0,935
Red neuronal de base radial	44,2 %	97,9 %	91,7 %	0,915
Regresión nominal	54,4 %	96,8 %	91,4 %	0,958
Regresión logística	52,9 %	97,0 %	91,4 %	0,919
Red neuronal perceptrón	29,1 %	97,9 %	90,0 %	0,898
Árbol QUEST	47,1 %	95,9 %	89,8 %	0,871
Promedio	51,64 %	96,68 %	91,18 %	91,35 %

Fuente: Elaboración propia

Como se puede apreciar, en la medida que el área bajo la curva ROC se sitúa en torno al 90 % en todos los casos, se puede afirmar que no existen diferencias estadísticamente significativas en el área de los distintos modelos considerados. Ahora bien, si nos centráramos en el porcentaje de acierto o precisión obtenido en la categoría "SÍ", solo sería aceptable la del árbol CHAID, con el que conseguimos catalogar correctamente a tres de cada cuatro entidades con problemas. En la clase "NO", por su parte, los resultados obtenidos con los distintos modelos son muy similares, con una precisión en torno al 96,7 %.

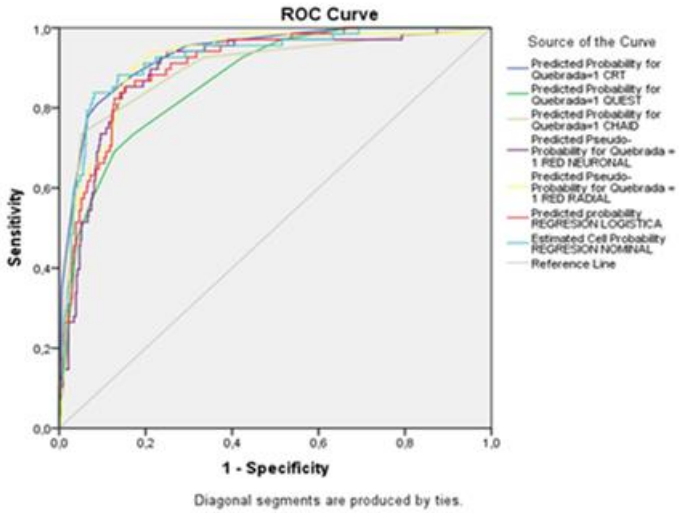

Gráfico 4. Curva ROC.
Fuente: Elaboración propia

3.2.3 Análisis y discusión de los resultados

De todos los modelos utilizados en este trabajo, teniendo en cuenta como criterio el área bajo la curva ROC (véase tabla 5), no se puede concluir que, estadísticamente, un modelo sea mejor que el resto. Efectivamente, en la medida que los valores del área bajo la curva ROC se concentran en torno a 0,9, se puede afirmar que el ajuste es adecuado en todos los casos, si bien, tal y como se desprende de la columna titulada "Total (%)", de la tabla 5, el modelo CHAID es el que obtiene la mayor precisión total (92,2 %), al igual que en la categoría "Sí", que es la relevante a los efectos de este trabajo. Adicionalmente, tal y como se tendrá ocasión de comprobar más adelante, se ha tenido en cuenta la sencillez en su interpretación y, por tanto, se ha decidido finalmente considerar este método como el más apropiado.

3.2.3.1 Árbol de decisión CHAID.

El diagrama de árbol de decisión CHAID (en su versión exhaustiva) obtenido en este trabajo se muestra en la figura 1. En ella se puede observar que las variables independientes seleccionadas por el algoritmo son, por su mayor poder explicativo, el saldo de la balanza corriente (Nodos 1, 2 y 3), el factor solvencia (Nodos 4, 5, 6 y 7) y el tamaño (Nodos 8 y 9). En el Nodo 0 se representan todos los casos -entidades con problemas (SI; 68 entidades) y sin problemas (NO; 469 entidades)-. Las observaciones (casos) se distribuyen por nodos teniendo en cuenta la significación de las variables dentro del modelo según el estadístico Chi-Cuadrado o F de Snedecor, según que las variables sean cualitativas o cuantitativas, de tal manera que los nodos terminales contienen el conjunto total de los casos (537), constituyendo particiones definidas según su capacidad predictiva.

Los resultados obtenidos con el árbol CHAID (en su versión exhaustiva) permiten concluir que el saldo de la balanza corriente es el mejor predictor para las entidades que han experimentado problemas. También se observa que cuando el saldo de la balanza por cuenta corriente tuvo un déficit superior a 4.758 millones de euros, solamente un 1,6 % de las entidades tuvo problemas (Nodo 1. Terminal). El exceso de crédito generado en los períodos expansivos pudo hacer que se relajasen los criterios de riesgo, con el peligro de generar inflación, perjudicando a las exportaciones, pero beneficiando a las importaciones de bienes de capital, necesarias para aumentar la producción, que tuvieron que ser financiadas.

Para las categorías del saldo de la balanza comercial con un déficit entre 4.758 y 3.226 millones de euros (Nodo 2) y el saldo de la balanza comercial con un déficit menor a 3.226 millones de euros (Nodo 3), el siguiente mejor predictor es la solvencia. Para valores de solvencia menores a -0,55 el modelo incluye un predictor más, el tamaño. Un 78,3 % de las entidades experimentaron problemas para valores de déficit de la balanza comercial superiores a 3.226 millones de euros, con valores de solvencia menores a -0,55 y con un tamaño menor a 0,44 (entidades financieras de mayor tamaño).

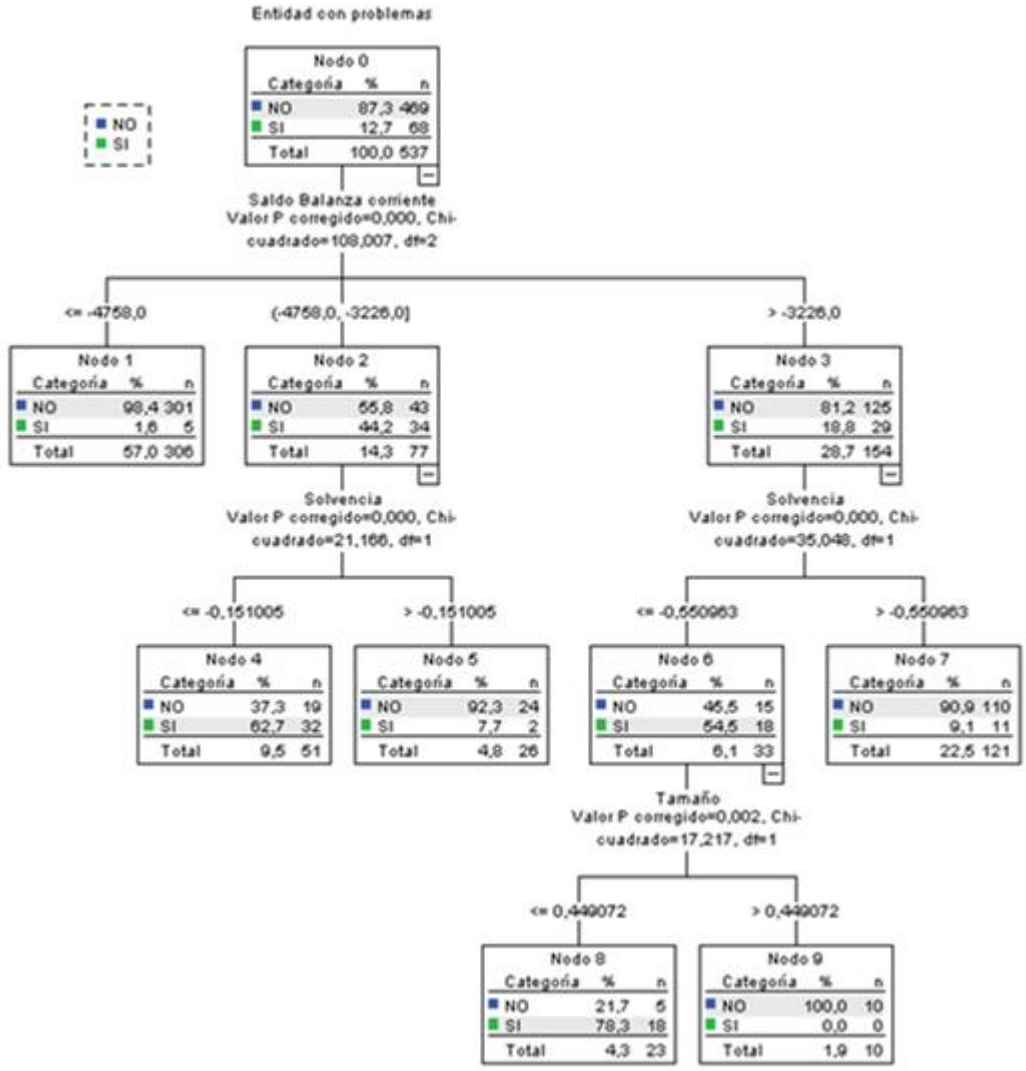

Figura 1. Árbol de decisión CHAID exhaustivo.
Fuente: Elaboración propia

Se observa que una de las probabilidades más altas de que las entidades no presenten problemas se produce en el Nodo 1, con el 98,4 %, cuando el déficit del saldo de la balanza comercial fue superior a 4.758 millones de euros. Por otra parte, se observa que la probabilidad más alta de que las entidades presentaran problemas se obtiene en el Nodo 8, con una probabilidad del 78,3 %. Esto se produce cuando el déficit de la balanza comercial disminuyó por debajo de 3.226 millones de euros, la solvencia de las entidades bajó de -0,55 y el tamaño de 0,44.

En la tabla 6 se recogen los nodos terminales. El nodo 8, por ejemplo, está formado por 23 casos, que suponen el 4,3 % del total, dieciocho de los cuales son entidades con problemas que representan un 26,5 % del total de las entidades con problemas (68). En la variable respuesta se observa el porcentaje de casos pertenecientes a la variable objetivo

en el nodo, donde en el caso del nodo 8 el 78,3 % de los casos son entidades con problemas. Por último, el índice es la proporción de respuestas de la variable objetivo en el nodo, en comparación con el porcentaje global de respuestas de la variable objetivo para toda la muestra, siendo para el nodo 8 el 618

Se observa cómo en los nodos 8 y 4 se concentra el 26,5 % y el 47,1 %, respectivamente, lo que representa el 73,6 % de entidades con problemas, frente al nodo 7, por ejemplo, en el que las entidades con problemas representan solo el 16,2 % del total de entidades con problemas, o los nodos 5, 1 y 9, que sólo concentran el 10,3 %.

Como se puede apreciar, cuando el saldo de la balanza comercial arrojó valores superiores a un déficit de 3.226 millones de euros y la solvencia presentó valores inferiores a -0,55, con un tamaño mayor, se produjo un 26,5 % de las quiebras de la muestra. Para un saldo de la balanza comercial peor (de entre - 4758 y - 3226 millones de déficit), con valores de solvencia mejores que los anteriores, pero inferiores a -0,15, se produjo un 47,1 % de las quiebras.

Tabla 6. Ganancias para los nodos CATEGORÍA OBJETIVO: SI

Nodo	Nodo		Ganancia		Respuesta	Índice
	N	Porcentaje	N	Porcentaje		
8	23	4,3 %	18	26,5 %	78,3 %	618,0 %
4	51	9,5 %	32	47,1 %	62,7 %	495,5 %
7	121	22,5 %	11	16,2 %	9,1 %	71,8 %
5	26	4,8 %	2	2,9 %	7,7 %	60,7 %
1	306	57,0 %	5	7,4 %	1,6 %	12,9 %
9	10	1,9 %	0	0,0 %	0,0 %	0,0 %

Métodos de crecimiento: CHAID exhaustivo.
Variable dependiente: Entidad con problemas
Fuente: Elaboración propia.

En la tabla 7, por su parte, se recoge tanto el riesgo como la clasificación de las entidades obtenida finalmente.

Tabla 7. Riesgo y clasificación

Riesgo	
Estimación	Típ. Error
,078	,012

Clasificación			
Observado	Pronosticado		
	NO	SI	Porcentaje correcto
NO	445	24	94,9 %
SI	18	50	73,5 %
Porcentaje global	96,1 %	32,4 %	92,2 %

Métodos de crecimiento: CHAID exhaustivo.
Variable dependiente: Entidad con problemas.
Fuente: Elaboración propia.

El riesgo es una medida de la capacidad predictiva del modelo. La parte de la tabla titulada "clasificación" muestra el número de casos clasificados correctamente e incorrectamente, para cada categoría de la variable dependiente. La estimación del riesgo, del 0,078, indica que la categoría pronosticada en el modelo de entidades con problemas y sin problemas es errónea para el 7,8 % de los casos. En consonancia con el riesgo, el modelo clasifica de forma correcta el 92,2 % de los casos. En concreto, las entidades que no experimentan problemas son pronosticadas correctamente en un 94,9 % de los casos, mientras que las que sí, se pronostican correctamente en un 73,5 %.

Por último, el gráfico 5 muestra la capacidad discriminatoria del modelo, a través del índice de ganancia de los nodos terminales. En un modelo que no proporcione información útil, la línea se situará en torno al 100 % a lo largo de todo el gráfico, mientras que un modelo útil presentará una línea con una pendiente decreciente significativa.

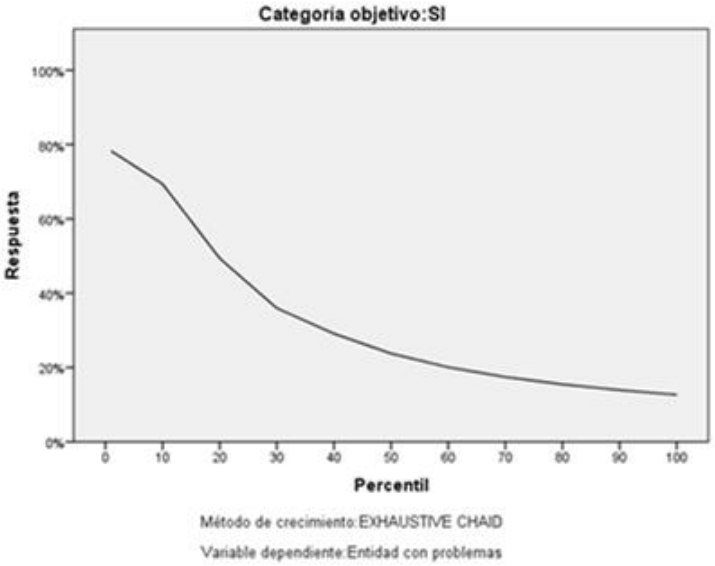

Gráfico 5.
Fuente: Elaboración propia.

4. Conclusiones

A pesar de la evidencia empírica previa obtenida en materia de selección de modelos de predicción de riesgos, en este trabajo se comprueba que el árbol CHAID en su versión exhaustiva es el mejor modelo de identificación de entidades de crédito (bancos y cajas de ahorros), en el sistema financiero español, con problemas de liquidez y solvencia, teniendo en cuenta tanto la situación económico-financiera en que se encontraba cada una ellas, en el período objeto de estudio (años 2005 a 2012), como la propia de la economía española.

El Data Mining puede ser visto, bien como el conjunto de estrategias que exploran las bases de datos, o bien como la forma en que se busca, entre los datos, los patrones o algoritmos que dan sentido a los comportamientos. Para ello, se sirve de herramientas procedentes tanto de la Estadística como del Machine Learning. Es evidente que la combinación de modelos, así como la utilización de diferentes algoritmos de cara modelizar cualquier tipo de suceso suele comportar una mejora significativa en la capacidad predictiva, y todo lo que ello conlleva. El principal inconveniente asociado al Data Mining, dado que suele estar asociado a grandes volúmenes de datos, es que puede requerir una enorme tarea y, por tanto, una mayor inversión en tecnología y un personal de mayor capacitación. Pero no todos los trabajos en que se aplica son de gran envergadura. En este sentido, los algoritmos más sencillos provenientes del Machine Learning han demostrado una gran utilidad en problemas que, por su volumen de datos, son abordados de manera habitual a través de técnicas estadísticas tradicionales, como los modelos logísticos, especialmente en aquellas ocasiones en que no se cumplen las hipótesis establecidas para los datos de cara a su aplicación. En situaciones como la que se presenta en este trabajo aportan, además, una interpretación más sencilla e intuitiva.

Si bien, en principio se podría afirmar que los resultados obtenidos por los distintos modelos testados (árbol CRT, árbol QUEST, perceptrón multicapa, funciones de base radial, regresión logística y regresión nominal) pueden calificarse como equivalentes en términos de capacidad predictiva, de acuerdo con los resultados obtenidos con la curva ROC, según la cual el área bajo dicha curva se sitúa en torno al 90 % en todos los modelos, sin embargo, hemos llegado a concluir que el árbol de decisión CHAID es el mejor, por cuanto que el porcentaje de acierto o precisión obtenido para la categoría "SÍ" (entidades con problemas), que es la relevante a efectos de este trabajo, es del73,5 %, muy por

encima del obtenido por el resto de modelos, que implica una capacidad del modelo para catalogar correctamente a tres de cada cuatro entidades con problemas. En la clase "NO" los resultados obtenidos con los distintos modelos son muy similares, con una precisión en torno al 96,7 %. Además de los resultados indicados, se debe recordar que este árbol presenta una mayor simplicidad frente al resto de modelos, al utilizar únicamente tres variables predictoras, dos microeconómicas (solvencia y tamaño) y una macroeconómica (saldo de la balanza de pagos). En consecuencia, el árbol de decisión CHAID podría ser fácilmente utilizado por el regulador español a efectos de identificación de entidades con posibles riesgos.

De las catorce variables microeconómicas definidas y calculadas a partir de la información publicada en los informes financieros de todos los bancos y cajas de ahorro que integran el sistema financiero español, para el período 2005 a 2012 (ambos inclusive), se obtienen finalmente, cuatro factores subyacentes. De ellos, se obtiene que el factor solvencia es el indicador microeconómico más relevante a efectos de identificación de entidades con problemas, entendiendo por tales aquellas que, o bien hubieran participado en procedimientos de recapitalización vía fusiones y adquisiciones o por aumentos de capital, o bien por haberse beneficiado de sistemas institucionales de protección como la concesión de ayudas públicas o, directamente, mediante la intervención de los poderes públicos. De las variables macroeconómicas empleadas, el saldo de la balanza de pagos es el primer predictor identificado por el modelo, indicando que un déficit acentuado de la balanza corriente genera entrada de capitales con la que financiar dicho déficit que, a su vez, provoca una expansión del crédito bancario de las entidades (mayor apalancamiento), con el consiguiente deterioro de su solvencia, al tiempo que genera mayor inflación y una menor competitividad.

De acuerdo con lo indicado, se puede concluir que cuando el saldo de la balanza por cuenta corriente presenta un déficit menor a -3.222,6 millones de euros, y el factor solvencia toma valores menores a -0,55, existe una probabilidad del 54,5 % de que una entidad bancaria española (banco o caja de ahorros) experimente problemas en el sentido identificado en este trabajo. En estas circunstancias, si se incorporan los efectos del factor tamaño, siendo este menor de 0,44 (mayor tamaño), se llega a la conclusión de que la probabilidad de que una entidad presente problemas bancarios sube hasta el 78,3 %, lo que podría interpretarse en el sentido de que nos encontramos ante un factor representativo de la conocida expresión "demasiado grande para quebrar", que sugiere la existencia de un apoyo implícito por parte del Estado español, que pudo llevar a estas entidades a asumir mayores riesgos, a pesar de no disfrutar de una buena solvencia, frente a entidades de tamaño mediano y pequeño, que sobrevivieron con una probabilidad del 100

Por último, los resultados de este trabajo nos llevan a concluir que en etapas de expansión crediticia, con un crecimiento del déficit de la balanza por cuenta corriente y de los riesgos, la probabilidad de que las entidades tengan problemas de solvencia es del 1,4 %, si bien, los riesgos acumulados se manifestarán en etapas posteriores de recesión, afectando a la solvencia de las entidades, que no podrán asumir dichos riesgos.

Referencias

Aktan, S. (2011) "Application of machine learning algorithms for business failure prediction". *Investment Management and Financial Innovations*, 8 (2), 52-65.

Alaka AH, Oyedeleb LO, Owolabi HA, Kumar V, Ajayi SO, Akinade OO, Bilal M (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems Applications*, 94:164-184.

Altman, E., R. Haldeman and P. Narayanan (1977), Zeta analysis: a new model to identify bankruptcy risk of corporations, *Journal of banking and finance*, p. 29-54.

Betz, F., Oprică S., Peltonen, A., Sarlin, P. (2013) "Predicting distress in european banks". *European Central Bank*, Working Paper Series, no. 1597/ october .

Boyacioglu, M. A., Kara, Y., Baykan, Ö. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications*(36), 3355–3366.

Brezigar-Masten, A., Masten, I. (2012). CART-based selection of bankruptcy predictors for the logit model. *Expert Systems with Applications*(39), 10153–10159.

Canbas, S., Cabuk, A., Bilgin Kilic, S. (2005). Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case. *European Journal of Operational Research*(166), 528–546.

Chen, M. Y. (2011)"Predicting corporate financial distress based on integration of decision tree classification and logistic regression". *Expert Systems with Applications*, 38 , 11261–11272.

Corona, E., Bejarano, V., González, J.R. (2017). *Análisis de estados financieros individuales y consolidados*. Ed. UNED. Madrid.

Cox, R., Wang, G. (2014). Predicting the US bank failure: A discriminant analysis. *Economic Analysis and Policy*(44), 202-211.

De Andrés, J., Lorca, P., de Cos Juez, F. J., Sánchez-Lasheras, F. (2011). Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). *Expert Systems with Applications*(38), 1866–1875.

Gepp, A., Kumar, K. (2015) "Predicting financial fistress: a comparison of survival analysis and decision tree techniques". *Procedia Computer Science*, 54, 396 – 404.

Irimía-Dieguez, A.I., Blanco-Oliver, A, Vázquez-Cueto, M.J. (2015) "A comparison of classification/-regression trees and logistic regression in failure models". *Procedia Economics and Finance*, 51, 396-404.

Kaiser, H. F. (1958) "The varimax criterion for analytic rotation in factor analysis". *Psychometrika*, 23 (3), 187-200.

Koyuncugil, A.S., Ozgulbas, N. (2009) "Risk modeling by CHAID decision tree algorithm". *ICCES*, 11(2), 39-46.

Le, H., Viviani, J.-L. (2017). Predicting bank failure: An improvement by implementing machine learning approach on classical financial ratios. *Research in International Business and Finance*.

Ley 11/2015, de 18 de junio, de recuperación y resolución de entidades de crédito y empresas de servicios de inversión.

Liviu, T., Mădălina, P., Marin, A. (2015) "A decision support system to predict financial distress. The case of Romania". *Romanian Journal of Economic Forecasting*, 18 (4), 170-179.

Directiva 2014/59/UE, de 15 de mayo de 2014, por la que se establece un marco para la reestructuración y la resolución de entidades de crédito y empresas de servicios de inversión, y por la que se modifican la Directiva 82/891/CEE del Consejo, y las Directivas 2001/24/CE, 2002/47/CE, 2004/25/CE, 2005/56/CE, 2007/36/CE, 2011/35/UE, 2012/30/UE y 2013/36/UE, y los Reglamentos (UE) no 1093/2010 y (UE) no 648/2012 del Parlamento Europeo y del Consejo (Directiva 2014/59/UE)

Madireddi, V. and Vadlamani, R. (2011). "Bankruptcy prediction in banks by principal component analysis threshold accepting trained wavelet neural network hybrid". *Proceedings of the 2011 International Conference on Data Mining*, 71-76.

Olson, D. L., Delen, D., Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support* Systems(52), 464–473.

Serhan Koyuncugil, A., Ozgulbas, N. (2012). Financial early warning system model anddata mining application for risk detection. *Expert Systems with Applications*(39), 6238–6253.

Serrano Cinca, C., Martín del Brío, B. (1993). Prediccion de la quiebra bancaria mediante el empleo de redes neuronales artificiales. *Revista Española de Financiación y Contabilidad*, 23(74), 153-176.

Sinkey F. Joseph(1975), A multivariate statistical analysis of the characteristics of problem banks. *The Journal of Finance*. Vol. 30, No. 1

Yiqiang Jin, J., Kanagaretnam, K., Lobo, G. (2011). Ability of accounting and audit quality variables to predict bank failure. *Journal of Banking Finance*(35), 2811–2819.

Anexo

Tabla 1. Ratios por años cajas de ahorros

Año	Ratio	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Media	,077	,084	,068	,862	,787	,717	1,003	,109	,009	,010	,030	6,982	5,460	,563
2005	**Desv. típ.**	,026	,032	,025	,053	,067	,123	,119	,063	,003	,003	,007	,498	,469	,073
	N	46	46	46	46	46	46	46	46	46	46	46	46	46	46
	Media	,073	,080	,064	,876	,808	,736	,958	,102	,008	,010	,029	7,031	5,548	,610
2006	**Desv. típ.**	,026	,031	,024	,041	,060	,120	,111	,057	,003	,003	,006	,547	,523	,076
	N	47	47	47	47	47	47	47	47	47	47	47	47	47	47
	Media	,070	,076	,063	,884	,797	,743	,957	,115	,008	,009	,028	7,094	5,718	,675
2007	**Desv. típ.**	,026	,032	,025	,040	,052	,119	,120	,051	,003	,003	,008	,562	,544	,074
	N	46	46	46	46	46	46	46	46	46	46	46	46	46	46
	Media	,060	,064	,055	,898	,760	,758	1,057	,115	,005	,004	,025	7,123	5,813	,723
2008	**Desv. típ.**	,022	,026	,021	,078	,102	,071	,160	,061	,006	,007	,006	,560	,549	,068
	N	46	46	46	46	46	46	46	46	46	46	46	46	46	46
	Media	,060	,064	,057	,898	,743	,725	1,091	,109	,002	,004	,026	7,134	5,713	,639
2009	**Desv. típ.**	,025	,029	,024	,081	,106	,076	,181	,053	,007	,006	,005	,559	,534	,080
	N	46	46	46	46	46	46	46	46	46	46	46	46	46	46
	Media	,053	,058	,050	,871	,722	,703	1,121	,103	-,001	,000	,016	7,433	5,511	,602
2010	**Desv. típ.**	,042	,046	,040	,146	,116	,098	,240	,053	,018	,014	,011	,732	,756	,263
	N	18	18	18	18	18	18	18	18	18	18	18	18	18	18
	Media	,057	,061	,054	,874	,714	,685	1,130	,101	-,001	,000	,020	7,565	6,030	,689
2011	**Desv. típ.**	,021	,024	,021	,108	,101	,112	,253	,045	,010	,010	,008	,683	,658	,082
	N	25	25	25	25	25	25	25	25	25	25	25	25	25	25
	Media	,033	,035	,032	,866	,617	,601	1,329	,115	-,040	-,032	,020	7,595	6,052	,617
2012	**Desv. típ.**	,034	,036	,033	,142	,143	,116	,245	,058	,050	,035	,005	,742	,714	,075
	N	20	20	20	20	20	20	20	20	20	20	20	20	20	20
	Media	,064	,069	,058	,881	,759	,720	1,051	,109	,002	,004	,026	7,172	5,701	,642
Total	**Desv. típ.**	,029	,033	,027	,081	,102	,111	,193	,056	,019	,015	,008	,612	,591	,109
	N	294	294	294	294	294	294	294	294	294	294	294	294	294	294

Tabla 2. Ratios por años bancos.

Año	Ratio	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Media	,102	,125	,102	,833	,819	,822	,935	,055	,008	,009	,045	6,588	5,062	,633
2005	**N**	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000
	Desv. típ.	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	Media	,105	,130	,104	,850	,834	,821	,908	,056	,012	,014	,048	6,651	5,189	,661
2006	**N**	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000
	Desv. típ.	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	Media	,108	,132	,102	,824	,828	,812	,936	,062	,011	,013	,049	6,714	5,366	,684
2007	**N**	29,000	29,000	29,000	29,000	29,000	29,000	29,000	29,000	29,000	29,000	29,000	29,000	29,000	29,000
	Desv. típ.	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	Media	,098	,117	,092	,866	,822	,818	,954	,068	,004	,005	,042	6,710	5,409	,900
2008	**N**	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000
	Desv. típ.	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	Media	,107	,136	,117	,856	,791	,788	,974	,081	,002	,005	,045	6,679	5,171	,885
2009	**N**	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000
	Desv. típ.	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	Media	,108	,136	,118	,853	,763	,757	1,057	,094	,001	,003	,044	6,668	5,047	,871
2010	**N**	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000
	Desv. típ.	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	Media	,103	,126	,106	,856	,725	,721	1,308	,102	-,001	,001	,043	6,636	5,082	1,039
2011	**N**	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000
	Desv. típ.	0	0	0	0	0	0	1	0	0	0	0	1	1	0
	Media	,097	,114	,096	,837	,688	,685	1,547	,114	-,008	-,003	,042	6,589	5,028	,844
2012	**N**	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000	31,000
	Desv. típ.	0	0	0	0	0	0	1	0	0	0	0	1	1	0
	Media	,103	,127	,105	,847	,784	,778	1,078	,079	,004	,006	,045	6,654	5,168	,814
Total	**N**	243,000	243,000	243,000	243,000	243,000	243,000	243,000	243,000	243,000	243,000	243,000	243,000	243,000	243,000
	Desv. típ.	0	0	0	0	0	0	1	0	0	0	0	1	1	0