PREDICCIÓN DE FRAGILIDAD FINANCIERA PARA SOCIEDADES ANÓNIMAS COLOMBIANAS MEDIANTE LA APLICACIÓN DE LAS TÉCNICAS LOGIT, ÁRBOLES DE CLASIFICACIÓN Y BOOSTING.

SUSANA ANDREA QUIRÓS HERNÁNDEZ

– ADMINISTRADORA FINANCIERA –

DIANA CATALINA REDONDO PANESSO

– PROFESIONAL EN FINANZAS Y COMERCIO INTERNACIONAL–

TRABAJO DE GRADO PARA OPTAR AL TITULO DE:

MAGISTER EN FINANZAS

Asesor:

MAURICIO LOPERA

Magister en Estadística

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE CIENCIAS ECONÓMICAS

MAESTRÍA EN FINANZAS

MEDELLÍN, ANTIOQUIA

2018

PREDICCIÓN DE FRAGILIDAD FINANCIERA PARA SOCIEDADES ANÓNIMAS COLOMBIANAS MEDIANTE LA APLICACIÓN DE LAS TÉCNICAS LOGIT, ÁRBOLES DE CLASIFICACIÓN Y BOOSTING

RESUMEN

Este trabajo presenta los resultados de aplicar tres diferentes técnicas para la predicción de quiebra empresarial para los tipos societarios de Sociedad Anónima (S.A. y S.A.S) en Colombia: regresión logística, árboles de clasificación y boosting. Inicia describiendo qué se entiende por quiebra empresarial y hace un recuento de los métodos de aprendizaje estadístico utilizados para su predicción, posteriormente se detalla el procedimiento metodológico utilizado, incluyendo los criterios de depuración y tratamiento de datos, así como el proceso de estimación de los modelos, para finalmente presentar el conjunto de indicadores financieros que explican en mayor medida la quiebra empresarial: solvencia y nivel de endeudamiento; así como cuál es el modelo de predicción más potente, haciendo uso de criterios como la tasa de clasificación y la métrica basada en el AUC (área bajo la curva ROC). El trabajo desarrollado desvirtúa el supuesto de que el modelo de regresión logística es insuficiente para predecir eventos raros de quiebra y señala algunas diferencias entre los modelos que pueden contribuir al refinamiento estadístico de las mediciones.

Palabras clave: regresión logística, árboles de clasificación, boosting, quiebra empresarial, indicadores financieros.

INTRODUCCIÓN

Los esfuerzos que se han realizado durante las últimas décadas en el estudio de la *quiebra empresarial*, designada también como *fracaso empresarial, fragilidad financiera, bancarrota o empresa fallida,* se enmarcan en la estadística y la minería de datos[1], como un problema de clasificación de individuos en grupos. En efecto, dado que el porcentaje de quiebra en una población no es muy significativo, la quiebra empresarial tiende a considerarse un evento raro, pero de gran importancia; ya que ésta, se presenta como un estado que no solamente compromete los intereses de la entidad y sus acreedores, sino en su totalidad a los stakeholders, demostrando la significancia que tiene su detección temprana para tomar medidas correctivas y evitar (o minimizar) sus efectos.

En vista de las consecuencias y a la incertidumbre asociada a esta situación, se han documentado, investigado y desarrollado una amplia gama de métodos que apuntan a la explicación y predicción de la quiebra. Dentro de éstos prevalece el desarrollo de modelos de elección binaria (Goleţ, 2014) los cuales discriminan entre empresas sanas y quebradas, asignando a la observación (empresa) una de esas dos categorías (James, Witten, Hastie, & Tibshirani, 2013).

Si bien no es posible saber con exactitud si una empresa se quiebra o no, si es posible predecir probabilísticamente su fragilidad financiera (Perez, Gonzalez, & Lopera, 2013); por ello, este trabajo se interesa en emplear tres métodos de predicción de quiebra para ser aplicados en el caso colombiano: regresión logística, árboles de clasificación y boosting. Estos métodos se utilizan para estimar la probabilidad de que un evento ocurra; es decir, buscan la probabilidad de que una observación pertenezca a un conjunto determinado en función de unas variables de entrada; de esta manera se valora la probabilidad de que una empresa pertenezca al grupo de empresas fracasadas o empresas no fracasadas (Mora, 1994).

La elección de estas tres metodologías se basa criterios disímiles: por un lado, la popularidad que para los analistas tiene la *regresión logística*, dada su facilidad de uso, por otro la consolidación de *árboles de clasificación* como una herramienta potente cuya interpretación es sencilla y ha mostrado ser efectiva para predecir eventos raros; y la eficacia teórica que tiene el

[1] Proceso de extracción de información fundamental de bases de datos grandes que permite tomar decisiones y aprender sobre el fenómeno estudiando, sin requerir de ningún conocimiento previo. Éste contiene una serie de técnicas, métodos y modelos que buscan descubrir las relaciones existentes entre un conjunto de datos, técnicas que surgen de la estadística, el aprendizaje automático, reconocimiento de patrones, entre otros (Bouza & Santiago, 2012).

boosting al ser una técnica basada en la combinación de una cantidad significativa de árboles. Ahora bien, indistintamente de la técnica utilizada, la probabilidad de quiebra se evalúa con razones financieras como variables de entrada, pues han demostrado ser, desde la década de los sesenta, una herramienta muy ventajosa para pronosticar el éxito o fracaso de una compañía, incluso años antes de su ocurrencia (Martínez, 2003).

Así, el objetivo de este estudio se centra en comparar y determinar cuál de los tres modelos es más potente, partiendo de la premisa teórica de que el boosting presenta mejor poder predictivo que los dos restantes, cuando de *eventos raros* se trata. Asimismo, a pesar de que a escala mundial existen múltiples estudios que desarrollan este tipo de modelos, para el caso de Colombia es marginal la aplicación de estas técnicas de medición y menos aún un esfuerzo por hacer un estudio comparativo, por lo que se busca actualizar un modelo de predicción de fragilidad financiera para las compañías colombianas, cuya razón social se limite a Sociedad Anónima (S.A) o a Sociedad por Acciones Simplificada (S.A.S), dados doce (12) indicadores financieros comúnmente utilizados dentro del diagnóstico en las organizaciones, calculados a partir de la información financiera reportada por la Superintendencia de Sociedades de Colombia en el año 2016. Para este propósito se cuenta con el apoyo del paquete estadístico R.

DEFINICIÓN DE QUIEBRA EMPRESARIAL

A pesar de la profusa literatura, acepciones y trabajos empírico – teóricos sobre el fracaso empresarial y el auge de modelos de predicción, no existe una única manera de comprender la quiebra; aunque tales desarrollos permiten identificarla con tres conceptos: dejar de pagar una deuda, reunir las condiciones previstas en la normativa vigente sobre quiebra (empresas legalmente en quiebra) y tener una situación patrimonial predecesora del fracaso futuro (Tascón & Castaño, 2012), lo que de entrada conduce a afirmar que no existe una forma “correcta” o “incorrecta” de identificar a una empresa como fracasada o no fracasada. Una síntesis de cómo se ha entendido este fenómeno se presenta en la siguiente tabla.

Tabla 1: Definiciones de quiebra empresarial utilizados por autores representativos

Autor	Término utilizado	Definición
Alman	Quiebra	Aquellas empresas que se encuentran legalmente en quiebra.
Beaver	Fracaso	La incapacidad de la empresa para atender sus obligaciones financieras a su vencimiento.
Blum	Fracaso	Incapacidad de pagar las deudas por parte de la empresa, entrando en un proceso de quiebra o en un acuerdo para reducir dichas deudas.
Deakin	Fracaso	Empresas que se encuentran en situación de quiebra, insolvencia, o fueron liquidadas a beneficio de los acreedores.
Taffler	Fracaso	Liquidación voluntaria, orden legal de liquidación o intervención estatal.
Zmijewski	Fracaso	Solicitar la quiebra

Fuente: Mora, E. A. (1994). pp. 712.

En Colombia, la situación de quiebra está asociada entre otros factores, al elevado endeudamiento, las altas tasas de interés, la reducción en las ventas, los malos manejos administrativos, la alta competencia y la falta de personal competente dentro de las organizaciones (Supersociedades, 2012). Estos factores, cuando se presentan de manera recurrente, pueden conducir a las organizaciones a un posible en estado de insolvencia, que se inicia desde el decreto de *disolución* de la empresa (que implica suspensión de sus actividades), hasta la etapa final cuando la sociedad es declarada legalmente en quiebra, pues no es viable una reorganización[2].

En Colombia se han establecido varias causales de disolución para cada uno de los tipos de sociedad, uno de ellos dado en términos de la reducción en el patrimonio o en el capital por debajo de un cierto porcentaje, lo que indicaría que la empresa estaría inmersa en una situación de quiebra técnica por lo que, salvo en caso de reorganización, su destino es la liquidación. En el caso de una *sociedad anónima,* el Código de Comercio Colombiano (Decreto 410 de 1971) determina que ésta se disolverá cuando el patrimonio neto esté por debajo del 50% del capital suscrito; y a su vez, la Ley 1258 de 2008, estipula que una *sociedad por acciones simplificada* (S.A.S) se disolverá cuando se den pérdidas que reduzcan el patrimonio neto de la sociedad por debajo del cincuenta por ciento del capital suscrito, al igual que una sociedad anónima. Es menester aclarar que para la

[2] Proceso nacido con la ley 1116 de 2006 que "pretende a través de un acuerdo, preservar empresas viables y normalizar sus relaciones comerciales y crediticias, mediante su reestructuración operacional, administrativa, de activos o pasivos" (Supersociedades, 2012, p. 29), con el fin de evitar una declaración legal de quiebra y liquidación judicial.

interpretación de estos criterios legales el punto de comparación es el capital suscrito, que incluye el capital suscrito por pagar, y no simplemente el capital efectivamente pagado

MODELOS DE FRAGILIDAD FINANCIERA EN EL MUNDO

Los primeros modelos aplicados para detectar la fragilidad financiera de una compañía fueron el modelo univariado aplicado por Beaver en 1966 y los modelos multivariados de Altman en 1968, trabajos pioneros que se caracterizaron por utilizar sólo indicadores financieros en su intento por predecir bancarrotas (Martínez, 2003); posteriormente surgieron modelos como el análisis discriminante (MDA), desde el cual se originó el método Z-Score por Altman en 1977, y los modelos Logit y Probit aplicados por Ohlson en 1980 (Ringeling, 2004).

Una evolución posterior dio origen a métodos más sofisticados y computacionalmente más intensivos; a partir de la década de los 90 se introduce en el estudio de la *quiebra empresarial* la aplicación de técnicas de inteligencia artificial, en especial metodologías basadas en redes neuronales y *árboles de decisión* (Tascón & Castaño, 2012), que se han vuelto cada vez más utilizados para estimar la probabilidad de ocurrencia de eventos raros, así como otras técnicas más modernas de clasificación que igualmente son bastante efectivas y capaces de arrojar predicciones precisas; por ejemplo, los conjuntos ásperos, algoritmos genéticos y el vector de máquinas. No obstante, parece ser que estos no garantizan mejores resultados (Alaminos, Del Castillo, & Fernández, 2016).

De todas maneras, no hay una conclusión definitiva de cuál de las metodologías es la más adecuada y más precisa para construir modelos. En la *tabla 2* se relacionan las diferentes metodologías aplicadas hasta la actualidad para la estimación de modelos predictivos de fragilidad financiera.

Tabla 2: Metodologías aplicadas en la predicción de quiebra empresarial hasta la actualidad

Fecha	Metodología
Años 30 del siglo XX	Modelos univarianes básicos
1966	***Beaver:*** Análisis univariante. Análisis de la varianza y el test de clasificación dicotómica.
1968	***Altman:*** Análisis discriminante Multivariado. Modelo Z-score.
1977	***Martin:*** Regresión logística. Logit y probit.
1984	***Marais et al.:*** Algoritmo de particiones recursivas o iterativas.
1990	***Bell et al.:*** Inteligencia artificial. Redes neuronales.
1991	***Mar Molinero & Ezzamel:*** Técnicas de escalamiento multidimensional.
1996	***Serrano-Cinca:*** Inteligencia artificial: mapas autoorganizativos.
2002	***Park & Han:*** Análisis multicriterio.
2002	***Shin y& Lee:*** Inteligencia artificial. Algoritmos genéticos.
2004	***Paradi et al.:*** Análisis Envolvente de datos (DEA)

Fuente: Mora, E. A. (1994). Limitaciones metodológicas de los trabajos empíricos sobre la predicción del fracaso empresarial. Revista española de financiación y contabilidad, 24(80), pp. 709-732.

Según un trabajo realizado por Alaminos, et. al (2016), la construcción de modelos para ofrecer predicciones estrictas de quiebra están centrados en su mayoría en un país o en una industria en particular; de estos estudios, los más importantes se presentan en la *tabla 3*, donde se observa que, efectivamente, la regresión logística y modelos de inteligencia artificial son preponderantes en este tipo de estudios.

Aun así, un estudio realizado por Adnan Aziz & Dar (2006), que analiza de manera crítica un gran número de estudios empíricos de la predicción de quiebra corporativa, concluye que las técnicas estadísticas, en particular los modelos MDA y Logit, se han utilizado con mayor frecuencia, a pesar de que los sistemas expertos inteligentes funcionan marginalmente mejor que los modelos estadísticos.

Tabla 3:Estudios aplicados en la construcción de modelos de predicción empresarial a nivel mundial

		Autor(es)	Año	Modelos utilizados	Precisión del modelo
	Estados unidos	Odom, M & Sharda, R.	1990	Modelo MDA.	86,80%
		Odom, M & Sharda, R.	1990	Redes neuronales.	77%
		Wilson, R & Sharda, R.	1994	Redes neuronales.	100% en fase de entrenamiento y 97% en fase de evaluación
		Mossman, C., Bell, G.,Swartz, L. & Turtle, H.	1998	Análisis discriminante y modelo de probabilidad lineal.	84,9%
		Laitinen EK, Laitinen T.	2000	regresión logística.	
		Shumway, T.	2001	Regresión logística.	54%
Europa	**Empresas eslovenas**	Brezigar-Masten, A & Masten, P.	2012	Árboles de decisión y regresión no paramétrica.	Entre el 65% y 95%
	Empresas de Reino Unido	Tinoco, M & Wilson, N.	2013	Regresión logística y redes neuronales.	
	Rusia	Fedorova, E., Gilenko, E. & Dovzhenko, S.	2013	Análisis discriminante multivariado, redes neuronales, árboles de decisión y regresión logística.	87,8%
	Empresas Francesas	DuJardin, P.	2015	Redes neuronales, modelo Cox y regresión logística.	81,2%
	Empresas Belgas	Cultrera, L. & Brédart, X.	2015	Regresión logística.	
Otras	**Empresas Coreanas**	Jo, H., Han, I. & Lee, H.	1997	Razonamiento basado en casos y Redes neuronales.	Entre 81,5% y 83,8%
	Empresas australianas	Hensher, D. & Jones, S.	2007	Regresión logística.	96%
	Empresas chinas	Li, S. & Wang, S.	2014	Regresión logística.	97,1% dentro de muestra y 94,1% por fuera de muestra.

Fuente: *elaboración propia basada en Alaminos et al. (2016)*

MODELOS DE FRAGILIDAD FINANCIERA EN COLOMBIA

En Colombia se encuentra una cantidad reducida de aplicaciones de modelos para predicción de quiebra, dentro de estos se destacan los estudios realizados por Rosillo (2002) y Martínez A. (2003),el primero realiza un trabajo titulado "Modelo de predicción de quiebras de las empresas colombianas" en el que plantea un modelo cuyo objetivo es determinar que indicadores financieros poseen mayor potencial predictivo para identificar quiebra o situación financiera difícil. Este modelo utiliza como base la técnica de análisis discriminante diseñado por Edward Altman en 1968, utilizando una muestra de 106 empresas y la aplicación de 12 indicadores sacados de los estados financieros reportados por éstas. El modelo resultante no sólo tuvo un nivel de acierto del 100% para aquellas que resultaron fuertes y del 82% para las débiles, sino que concluyó que el modelo si servía para predecir si una empresa es frágil o no a través de los indicadores financieros obtenidos mediante el análisis discriminante. Finalmente, se halló que los

indicadores más importantes para el diagnóstico y predicción de quiebra empresarial son *endeudamiento, rentabilidad y apalancamiento*.

Por su parte, Martínez (2003) estableció un modelo probit heteroscedástico que permitió identificar las variables más importantes para medir la fragilidad financiera de las empresas colombianas, a partir de un análisis de indicadores financieros de 9.000 empresas, vigiladas por la Superintendencia de Sociedades, que reportaron estados financieros en el año 2000. Durante el estudio, el autor encontró 171 empresas de la muestra en estado de fragilidad financiera, término que definió por la situación legal de la empresa, en cuanto a si comenzó un proceso de reestructuración de pagos o si la Superintendencia de Sociedades estipuló la liquidación.

Mediante los resultados obtenidos, se identificaron indicadores de *liquidez, rentabilidad y endeudamiento* como los más relevantes en el momento de medir el estrés financiero en Colombia en el año 2001, en específico las razones: *disponible sobre activos, utilidad antes de impuestos sobre activos y obligaciones financieras sobre activos*. De igual forma, el modelo determinó correctamente el 82% de la muestra como empresas frágiles y no frágiles, a partir de esos tres indicadores y variables dummies sectoriales. Adicionalmente, estableció que las empresas menos propensas a ser frágiles durante el 2001, independiente de sus estados financieros, fueron las que pertenecen a actividades auxiliares a la intermediación financiera y actividades inmobiliarias, empresariales y de alquiler, mientras que las más propensas pertenecen al sector "enseñanza, servicios de salud y otros servicios".

INDICADORES FINANCIEROS Y QUIEBRA EMPRESARIAL

Los ratios financieros son transformaciones de los datos de los estados financieros, para servir como base en a la toma de decisiones de la compañía. Por ello, se utilizan, entre otras cosas, para la predicción de la bancarrota. Éstos permiten una comparación directa entre empresas de diferentes tamaños, y establecer una mejor imagen de la posición financiera (Elam, 1975). El uso de los indicadores financieros para predecir la quiebra empresarial se ha presentado desde la aplicación de los primeros modelos de predicción (Wu, Gaunt, & Gray, 2010). A partir de entonces, han evidenciado muy buen desempeño y convertido en las variables más populares que se consideran en la literatura (Li, Crook, & Andreeva, 2014).

Beaver (1966) fue el primer autor que introdujo las razones financieras dentro del análisis de la predicción de la bancarrota. De ahí, otros investigadores como Altman (1993), Balcaen and Ooghe (2006), Kumar and Ravi (2007), Bahrammirzaee (2010), y otros más recientes han demostrado que los ratios son los que dominan la elección de las variables explicativas en las diferentes técnicas de aprendizaje estadístico (Li et al., 2014). Muchos de ellos llevan a la conclusión de que las razones financieras son útiles para predecir quiebras corporativas (Mulyawan, 2015), a pesar de que es ampliamente reconocido que la principal causa de la quiebra de una empresa es la pobre gestión de los administradores (Gestel et al., 2006)

A lo largo de los años se ha establecido que los ratios tienen un poder predictivo de hasta por lo menos cinco años antes de la bancarrota (Beaver, McNichols, & Rhie, 2005). Asimismo, se ha demostrado que son útiles en condiciones de alta incertidumbre como las crisis económicas, e incluso, que son beneficiosos para predecir el crecimiento de las ganancias y las quiebras bancarias (Mulyawan, 2015).

Un estudio realizado por Mulyawan (2015) determina que la precisión de los indicadores financieros en la predicción es cada vez mayor a medida en que se aproxima la quiebra. Se demuestra que cuatro años antes de que una empresa se declare en quiebra se manifiestan diferencias significativas entre los ratios financieros de una empresa en bancarrota con respecto a una sana, pues las proporciones de los indicadores de liquidez, rentabilidad, actividad y rendimiento de la compañía sana son más altas, mientras que las de una empresa en quiebra son mucho menores. Además, se estableció que la liquidez y la solvencia son los ratios financieros dominantes en la predicción de quiebra.

De igual forma, otro estudio realizado por Contreras (2016) con una muestra de todas las empresas cotizadas de México, IBEX-35 de España y EURO STOXX50 de Europa en un horizonte temporal de 5 años, muestra que las variables más significativas en cuanto a predicción de quiebra son los ratios financieros, en concreto el ratio coeficiente de liquidez circulante, el ratio de solvencia global y el ratio de rentabilidad económica (ROA).

Finalmente, cabe la pena destacar que el desarrollo de los estándares de contabilidad, cuya intención es hacer que los estados financieros sean más útiles para inversionistas y otro tipo de usuarios, ha demostrado mejorar la capacidad de predicción de bancarrota con las razones financieros. (Beaver et al., 2005).

REGRESIÓN LOGÍSTICA, ÁRBOLES DE CLASIFICACIÓN Y BOOSTING

La regresión logística es una metodología que, debido a que es fácil de utilizar y sencilla de interpretar, ha sido frecuentemente comparada, por muchos investigadores, con otras técnicas de predicción, entre ellas varias pertenecientes al análisis discriminante, resultando en su mayoría conclusiones a favor de la regresión logística (Serna, 2009). Ésta en vez de modelar una respuesta directamente para la variable dependiente quiebra (Y), modela la probabilidad de que pertenezca a una categoría en particular, el de empresa frágil o no frágil, dado los valores de las variables explicativas. Los valores de probabilidad (valor de la variable independiente) están en un rango de entre 0 y 1, cuando toma el valor de 0 significa que la empresa no ha fracasado y cuando toma el valor de 1 simboliza que ha fracasado. La respuesta de la variable dependiente es binaria y sus covariables pueden ser de carácter tanto categóricas como continuas (James et al., 2013).

No obstante, este procedimiento presenta una algunos inconvenientes cuando se emplean en eventos raros, dentro de los que destacan la subestimación aguda de la probabilidad y el sesgo amplificado de los parámetros estimados por la función de máxima verosimilitud debido a los tamaños generalmente pequeños de las muestras (King & Zeng, 2001). Conjuntamente, según Balcaen & Ooghe (2006), este tipo de modelos son muy sensibles a la multicolinealidad, a la existencia de valores extremos y a la falta de datos desaparecidos (Citado por Tascón F. & Castaño G., 2012).

De otro lado, la técnica de árboles de clasificación es un método no paramétrico[3] que aparece en la década de los 80 (fue propuesto por Leo Breiman en 1984) como un algoritmo de partición binaria que divide repetidamente la muestra de datos en grupos mutuamente excluyentes (Serna, 2009) y, a diferencia de los árboles de regresión, se utilizan con el objetivo de predecir un evento. Una de las grandes ventajas de esta metodología es que la interpretación del modelo final es fácil ya que se presenta de manera gráfica, por lo que cualquier persona estaría en la capacidad de comprenderlo, además de que elimina el supuesto de normalidad que involucran muchas metodologías paramétricas. Igualmente, el algoritmo que utiliza se adapta bien al número de datos de entrenamiento (datos que se utilizan para ajustar el modelo) y a un alto número de atributos de

[3] Métodos que suponen que la distribución proveniente de la muestra no está especificada, por ello se elimina el supuesto de normalidad que involucran muchas de las metodologías paramétricas.

las grandes bases de datos; asimismo, no está sujeta a restricciones como la homoscedasticidad[4], como lo está el modelo Logit.

La construcción de un árbol de esta índole consiste fundamentalmente en dos pasos: construcción del árbol inicial y poda. Utilizando datos de entrenamiento, el árbol inicial es construido mediante un procedimiento de partición binaria comenzando desde la raíz del árbol (datos sin ninguna división) la segmentación en dos nodos internos (también llamados nodos hijos) cada vez, y a su vez la partición es aplicada a cada nodo interno por separado. Al final, cada grupo es caracterizado por una respuesta categórica de la variable dependiente (dada por la moda de las regiones), el tamaño del grupo y los valores de las variables explicativas que lo definen (Serna, 2009). Debido a que el algunas veces, el resultado puede ser un árbol muy extenso y demasiado complejo, el árbol resultante inicialmente, es podado cortando sucesivamente ramas y nodos terminales hasta encontrar el tamaño adecuado del árbol. Mediante este proceso se busca alcanzar el máximo grado de pureza posible usando el menor número de particiones, de manera que el árbol resultante sea pequeño y el número de datos de cada subconjunto grande (Diaz, 2000).

Pese a la facilidad y simpleza que presenta este método en cuanto a la interpretación y entendimiento, su principal problema radica en que, a pesar de presentar resultados satisfactorios, no son tan competitivos en términos de precisión de predicción cuando se le compara con otros enfoques de regresión y clasificación. No obstante, el rendimiento predictivo este tipo de modelos se pueden perfeccionar sustancialmente con técnicas basadas en la combinación de una cantidad significativa de árboles, para posteriormente producir en consenso una predicción única (Seoane, Carmona, Tarjuelo, & Planillo, 2014). Éstas han evidenciado presentar mejoras dramáticas en la precisión del modelo, a expensas de algunas pérdidas en la interpretación (James et al., 2013). Dentro de este tipo de técnicas, se encuentran el bagging, random forest y *boosting*, estrategias basadas en la mezcla de modelos; en este caso, se generan y combinan varios árboles en vez de uno solo, para mejorar la precisión de un modelo individual basado en árboles.

Cada una de esas técnicas establece un procedimiento para construir diferentes modelos predictores y una regla para combinar los resultados. En el caso del boosting, busca crear varios modelos secuencialmente de manera adaptativa, de manera que asignando un peso a cada modelo,

[4] Característica que supone que un modelo presenta varianza de los errores de forma constante a lo largo del tiempo. Es decir, los sujetos o datos se comportan de igual forma, lo cual supone una serie de hipótesis básicas sobre la distribución de los datos.

se obtiene un modelo final más eficiente (Bourel, 2012). El algoritmo AdaBoost es el algoritmo del tipo boosting más conocido, desarrollado por Freund y Schapire en 1997, pensado originalmente para la clasificación binaria. El primer paso es crear un único árbol, el cual seguramente cometerá algunos errores de clasificación, los cuales serán el foco de atención al construir el segundo árbol, que será diferente del primero y también cometerá errores en la clasificación, que también serán el foco de atención en la construcción del tercer modelo (Diaz, Fernandez, & Segovia, 2004). Este proceso se realiza de manera repetitiva de manera que se obtiene un modelo realmente preciso que disminuye la varianza y el sesgo (Bourel, 2012).

La forma en cómo se tienen en cuenta los datos mal clasificados del árbol anterior, se da mediante la manipulación los pesos de los datos en el conjunto de entrenamiento. Inicialmente, todos los objetos del conjunto poseen el mismo peso, pero cada peso irá evolucionando a medida que avanza el algoritmo; es decir, en cada iteración se incrementa el peso de cada uno de los objetos mal clasificados por el modelo en esa iteración, por lo tanto, en la construcción del próximo modelo, estos objetos serán más significativos (Bourel, 2012).

En conclusión, el boosting va ajustando modelos de forma secuencial de manera que en cada paso va poniendo más énfasis en las observaciones que no se han ajustado bien en árboles anteriores (Seoane et al., 2014). Finalmente, los árboles utilizados con esta técnica suelen ser pequeños y no interesa que el árbol se ajuste en demasía a las características de los datos sino que el procedimiento vaya aprendiendo lentamente (Seoane et al., 2014).

METODOLOGÍA

Muestra y depuración de los datos

La información empleada ha sido obtenida de los estados financieros de 15.524 empresas vigiladas por la Superintendencia de Sociedades de Colombia y extraídos de la base de datos reportada por dicha institución al 31 de diciembre de 2016 en su Sistema de Información Empresarial (SIE). Inicialmente, la base de datos estaba compuesta por 22.054 empresas; sin embargo, dentro del tratamiento de los datos, se presentaron algunos inconvenientes relacionados con inconsistencias en los datos, ausencia de información y resultados indeterminados en las operaciones que no permitían su correcta utilización para propósitos del trabajo; por lo que, en

primera instancia fue necesario depurar la base de datos de manera tal que: la misma cantidad y las mismas empresas tuvieran información reportada para cada uno de los estados financieros, el tipo de sociedad de las empresas de la base de datos solo fuera anónima (S.A) y por acciones simplificadas (S.A.S) y, se hayan eliminado de la lista las empresas cuyos cálculos en uno o más indicadores financieros resultan matemáticamente indeterminados.

Clasificación de las empresas (variable dependiente): frágiles y no frágiles

Una vez depurada la información y obtenida una base de datos sólida, se procede a la clasificación de las empresas en dos grupos: frágiles o no frágiles. La anterior categorización se realiza teniendo en cuenta el concepto de causal de disolución de una sociedad que emite el Decreto 410 de 1971 (Código de comercio colombiano) en el artículo 457 y la Ley 1258 de 2008 en su artículo 34. La primera de ellas para una empresa cuyo tipo de sociedad es anónima y la segunda para una empresa por acciones simplificada. En ambos casos, la disolución se efectúa cuando se produzcan pérdidas en la empresa que hagan al patrimonio neto reducirse por debajo del 50% del capital suscrito. En caso de que se cumpla el causal de disolución, la empresa se considera financieramente frágil, en caso contrario se considera no frágil. En la *tabla 4* se presentan el número de empresas clasificadas de acuerdo a esa definición:

$$Sea\ y_i = \begin{Bmatrix} 1, si\ \ hay\ \ \ fragilidad \\ 0, si\ no\ hay\ fragilidad \end{Bmatrix}\ Donde\ \rightarrow fragilidad = \frac{Patrimonio\ neto}{Capital\ suscrito} < 0.5$$

Tabla 4: Clasificación de las empresas: frágiles o no frágiles

Clasificación	Número de empresas
Empresas no frágiles	14.698
Empresas frágiles	826
Total empresas	**15.524**

Fuente: *elaboración propia en base a la información reportada por la Superintendencia de sociedades de Colombia a diciembre 31 de 2016.*

Selección de las variables explicativas

Las variables con las que se pretende predecir si una empresa es financieramente frágil o no, son 12 de las razones financieras más utilizadas dentro del diagnóstico financiero empresarial, pertenecientes a los diferentes grupos de indicadores: liquidez y solvencia, rentabilidad, endeudamiento (y apalancamiento) y actividad. En la *tabla 5* se describe cada uno de los indicadores utilizados para este propósito.

Tabla 5: Definición de los indicadores financieros utilizados como variables independientes en los modelos

Grupo	Indicador	Definición matemática	Definición teórica
Liquidez y solvencia	Razón corriente	$\frac{Activo\ corriente}{Pasivo\ corriente}$	Evalúa la capacidad de la empresa, a corto plazo, de afrontar los su compromisos también a corto plazo. Por lo tanto, entre mayor sea el valor de la razón mejor se considera el indicador (desde el punto de vista del acreedor), comenzando a considerarse bueno a partir de una relación 1:1. Relaciones menores a 1 indican que la empresa no tiene la capacidad suficiente para cubrir sus obligaciones de largo plazo.
	Prueba ácida	$\frac{Activo\ corriente - inventarios}{Pasivo\ corriente}$	Verifica la capacidad que tiene la empresa para cubrir sus pasivos corrientes con sus activos corrientes sin depender de la venta de su inventario; es decir, trata de establecer que pasaría si se pararan súbitamente las ventas y se tuvieran que cancelar todos los pasivos corrientes. Para algunos autores, se considera que una relación de 0,5 a 1 satisfactoria, mientras que para otros se considera adecuada desde 0,7.
	Razón de tesorería/efectivo	$\frac{Efectivo}{Pasivo\ corriente}$	A comparación de los dos indicadores anteriores, ésta es una prueba más rigurosa que mide la capacidad de una empresa de pagar sus pasivos corrientes con sus activos mas líquidos (caja y bancos). Un valor mayor a 0,1 puede ser calificada como buena.
	Índice de solvencia	$\frac{Activo\ total}{Pasivo\ total}$	Calcula la idoneidad de una empresa de cubrir sus todos sus pasivos con la totalidad de los activos. Al igual que en la razón corriente, un mayor resultado en el índice, mejor para el acreedor; así mismo, la relación considerada más beneficiosa es 1:1, mientras que una relación por debajo de 1 no es la deseada.
Rentabilidad	Margen neto	$\frac{Utilidad\ neta}{Ingresos\ operacionales}$	Indica el porcentaje de la utilidad neta correspondiente a las ventas (ingresos operacionales); es decir, por cada peso vendido, cuántos pesos se han generado de utilidad neta.
	Rentabilidad sobre el patrimonio (ROE)	$\frac{Utilidad\ neta}{patrimonio}$	Mide el rendimiento que los socios o dueños de la empresa han obtenido sobre su inversión; en otras palabras, mide cuantos pesos de utilidad se generan por cada $1 peso de patrimonio.
	Rentabilidad sobre los activos (ROA)	$\frac{Utilidad\ neta}{Activo\ total}$	Mide cuánto de la utilidad neta corresponde del activo total, lo que significaría la cantidad de utilidad (en pesos) generada por cada peso invertido en activos totales. En otras palabras, mide la capacidad del activo para generar utilidades independientemente de la forma en como se hayan financiado.
Endeudamiento	Nivel de endeudamiento	$\frac{Pasivo\ total}{Activo\ total}$	Determina que porcentaje de participación tienen los acreedores de la empresa dentro de ésta. Además, evalúa la capacidad futura de endeudamiento de la empresa
	Apalancamiento a corto plazo	$\frac{Pasivo\ corriente}{Patrimonio}$	Define el porcentaje en que los dueños de la empresa están comprometidos con los acreedores en el corto plazo.
	Apalancamiento total	$\frac{Pasivo\ total}{Patrimonio}$	Establece el grado de compromiso del patrimonio de los socios para con los acreedores dela empresa.
Actividad	Rotación cartera	$\frac{Ingresos\ operacionales}{Cuentas\ por\ cobrar}$	Mide en promedio cuántas veces las cuentas por cobrar se hacen líquidos durante un periodo determinado
	Rotación activo total	$\frac{Ingresos\ operacionales}{Total\ activos}$	Indica cuántas veces rotan los activos totales en el año; es decir, cuánto se generó en ventas por cada peso invertido en en activos totales.

Fuente: *Elaboración propia basada en las definiciones proporcionadas por Ortiz (1998) y Cruz, Villareal, & Rosillo (2001)*

Estimación de los modelos

De acuerdo a James et al. (2013), el modelo resultante no solo se debe comportar bien con datos de entrenamiento, sino que también debe hacerlo en con datos de evaluación; por lo que, para la estimación de los modelos se divide la muestra en dos partes: un conjunto de datos de entrenamiento (X_1,Y_1), $(X_2,Y_2),\ldots,(X_n,Y_n)$ que se utilizan para estimar los modelos y el cual está constituido por el 80% de los datos de la muestra total, y un conjunto de datos de evaluación empleados para valorar el desempeño de los modelos resultantes, compuesto por el 20% restante de los datos. Como se ilustra en la *tabla 6*, se tiene una base de 12.419 empresas como datos de entrenamiento y 3.105 como datos de evaluación.

Tabla 6: Conjunto de datos de entrenamiento y de evaluación

Clasificación	Empresas totales	Datos entrenamiento (80%)	Datos evaluación (20%)
Empresas no frágiles	14.698	11.758	2.940
Empresas frágiles	826	661	165
Total	15.524	**12.419**	**3.105**

Fuente: elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016.

A continuación, se describen los algoritmos utilizados para la estimación de cada uno de los modelos aplicados en el estudio:

Modelo Logit.

Dadas n empresas (Y_i , X_i), donde X_i es un vector con n razones financieras y Y_i es una variable independiente binomial la cual indica el grupo al cual pertenece cada compañía, la probabilidad condicional de pertenencia de cada empresa al grupo de frágiles es:

$$P_i = Pr(Y_i = 1) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_i X_i}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_i X_i}} \quad (1)$$

Se utiliza el método general de máxima verosimilitud para hallar el estimar de los coeficientes β_0, $\beta_1 \ldots \beta_i$ de tal manera que la maximicen:

$$\ell(\beta_0, \beta_1 \dots \beta_1) = \prod_{i:y_{i=1}} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \quad (2)$$

Arboles de clasificación.

El algoritmo aplicado de construcción del árbol presenta los siguientes pasos expuestos por James et al. (2013):

1. Dividir el espacio predictor usando un método de particionamiento binario, bajo el supuesto de que se tienen p variables explicativas $x_1, x_2, \dots., x_p$; entre J regiones distintas y no superpuestas, de tal forma que se encuentren regiones R_1,R_2…Rj que minimicen la medida de impureza dada por el "índice de Gini", fórmula por la expresión

$$G = \sum_{K=1}^{K} \hat{P}_{mk} (1 - \hat{P}_{mk}) \quad (3)$$

2. Aplicar el costo de complejidad del árbol para obtener una secuencia de los mejores subárboles, como una función de **α**:

$$R_\alpha(T) = R(T) + \boldsymbol{\alpha} \left| \boldsymbol{T} \right| \quad (4)$$

3. Usar la validación cruzada de K-pliegues para elegir **α**. Para eso se dividen los datos de entrenamiento en K grupos mutuamente excluyente y de aproximadamente igual tamaño.

i. Sacar aparte uno de esos conjuntos por vez y aplicar los pasos 1 y 2 en todos los pliegues menos en el escogido.
ii. Evaluar la tasa de error de clasificación en el pliegue dejado afuera, como una función de **α**.

4. Promediar los resultados para cada valor de **α** y se escoge un **α** que minimice la tasa de error.

$$CV(k) = \frac{1}{k}\sum_{j=1}^{k}(MSE_j) \qquad \textbf{(5)}$$

5. Finalmente, el árbol que corresponda al valor escogido de **α** es el árbol óptimo.

Boosting.

1. Asignar a todos los datos del conjunto de entrenamiento un mismo peso, $\boldsymbol{w_i = \frac{1}{n}}$, donde n=número de datos.

2. Estimar un árbol con *d* divisiones (*d* + 1 nodos terminales) con los datos de entrenamiento y calcular la tasa de error de clasificación e identificar los datos mal clasificados.

$$err = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(X_i))}{\sum_{i=1}^{N} w_i} \qquad (6)$$

$$\propto_m = Log(\frac{1 - err_m}{err_m}) \qquad (7)$$

3. Incrementar los pesos a los datos mal clasificados.

$$w_i \leftarrow w_i * \exp[\propto_m * I(y_i \neq G_m\big((x_i)\big), \qquad i = 1,2, \dots n \qquad (8)$$

4. Entrenar un nuevo modelo usando los datos con los pesos modificados.

5. Se repite el paso 3 y 4 un número de veces ya fijado, el cual representa el número de árboles a estimar.

6. El modelo final se constituye como la ponderación de los pesos de todos los modelos; es decir, la predicción estará dada por:

$$G(x) = sign\left[\sum_{m=1}^{N} \propto_m G_m(x)\right] \qquad (9)$$

RESULTADOS

A continuación, se detallan los resultados arrojados para cada una de las metodologías utilizadas, estimados por el paquete estadístico R.

Modelo logit

Aplicando el método de regresión logística, estimado asumiendo que la probabilidad de fragilidad o quiebra de una empresa está dada por la *ecuación* 1 se llega a un modelo que contempla las 12 variables explicativas establecidas. En la *tabla 7* se detallan los coeficientes del modelo de regresión logística resultante, mediante el método de máxima verosimilitud (*ecuación 2*).

Tabla 7:Estimación del modelo Logít

	Estimate	Std. Error	z value	Pr(>|z|)*
(Intercept)	-14,741	0,598886	-24,614	< 2E-16
Razón corriente (razcor)	-0,003025	0,020422	-0,148	0,8822
Prueba ácida (pruacida)	-0,006691	0,004470	-1,497	0,1345
Razón de efectivo (razef)	0,09372	0,124298	0,754	0,4509
Índice de solvencia (solv)	0,49625	0,063219	7,850	4,17e-15
Margen neto (maneto)	-0,01845	0,023410	-0,788	0,4307
Rentabilidad sobre el patrimonio (roe)	-0,13362	0,066843	-1,999	0,0456
Rentabilidad sobre el activo (roa)	-4,70014	0,489852	-9,595	< 2e-16
Nivel de endeudamiento (endeud)	13,836	0,608445	22,739	< 2e-16
Apalancamiento total (leveraget)	-0,040770	0,009109	-4,476	7,62e-06
Apalancamiento a corto plazo (leveragecp)	-0,0103810	0,013824	-0,751	0,4527
Rotación activo total (rotat)	0,064014	0,054581	1,173	0,2409
Rotación cartera (rotcar)	0,0009050	0,006109	0,148	0,8822

Fuente: *elaboración en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Para medir la significancia de las variables, analizada en el apartado de los resultados, se utiliza la prueba de hipótesis sobre los coeficientes estimados. Según (Gujarati & Porter, 2013), a la hipótesis planteada se le conoce como hipótesis nula, y se denota con el símbolo H_0, la cual suele probarse frente a una hipótesis alternativa, o hipótesis mantenida, denotada como H_1. En este caso, se utiliza la hipótesis nula "cero". La determinación de dicha asociación está dada por el siguiente razonamiento:

Sea:

$H_0: \beta i = 0$ → No hay relación entre X y Y

$H_1: \beta i \neq 0$ → Hay alguna relación entre X y Y.

- Si, P-value < $\boldsymbol{\alpha}$=0,05, se rechaza $H_0: \beta i = 0$; por lo tanto, Xi es estadísticamente significativa en el modelo.
- Si, P-value < $\boldsymbol{\alpha}$=0,10, se rechaza $H_0: \beta i = 0$; por lo tanto, Xi es marginalmente significativa en el modelo.
- Si, P-value ≥ $\boldsymbol{\alpha}$=0,10, no se rechaza $H_0: \beta i = 0$; por lo tanto, Xi no es estadísticamente significativa en el modelo.

Donde, Alpha ($\boldsymbol{\alpha}$) es el nivel de significancia.

En el modelo se observa que los indicadores de solvencia, rentabilidad sobre el activo (ROA), nivel de endeudamiento y el apalancamiento total rechazan la hipótesis nula con un nivel de significancia del 0,05, por lo que se constituyen como variables estadísticamente significativas en el modelo; mientras, el indicador de rentabilidad sobre el patrimonio (ROE) rechaza la hipótesis nula con un nivel de significancia del 0,10 lo que lo hace una variable marginalmente significativa en el modelo. Sin embargo, analizado la significancia de los ratios en un nuevo escenario donde se deja de considerar al ROE como una variable explicativa, se obtiene una mejora en cuanto a los resultados de predicción (predice de manera correcta una empresa más en comparación del modelo inicial), y las variables estadísticamente significativas siguen siendo exactamente las mismas.

Por lo tanto, se afirma que de las cinco variables mencionadas, se consideran como las más útiles para explicar el nivel de fragilidad financiera de las empresas colombianas a todas, excepto a la rentabilidad sobre el patrimonio.

Igualmente, se puede notar que los signos esperados de muchas de las variables son consecuentes con la lógica económica, pues los signos negativos de razón corriente, prueba ácida, margen neto, ROE y ROA, y el signo positivo del nivel de endeudamiento representa que mayores niveles de liquidez y rentabilidad, combinados con bajos niveles de deuda deberían reducir la probabilidad de quiebra de una empresa.

En el caso de las rotaciones y apalancamientos, se encuentran signos no tan congruentes con la lógica económica. Lo natural sería que, a mayores niveles de rotación y menor nivel de apalancamiento, menor fuera la probabilidad de quiebra, cosa contraria a lo que representan sus actuales signos. No obstante, si se revisan los datos estadísticos de las empresas, más exactamente la media de cada una de las variables presentadas en la *tabla 8*, encontramos que, de hecho, las empresas frágiles presentan mayores niveles de rotación que las empresas no frágiles, y menos comprometido su patrimonio con sus acreedores, por lo cual los signos estarían correctos.

Finalmente, en el caso de la solvencia y la razón de efectivo, su signo no es apoyado ni por la lógica económica ni por el análisis descriptivo, ya que las empresas frágiles presentan menores niveles de éstos, por tanto, su signo debería ser negativo.

Tabla 8: Media de las variables explicativas

	Media	
	No Frágiles	**Frágiles**
Razón corriente	3,331	2,346
Prueba ácida	2,479	1,135
Razón de efectivo	0,500	0,281
Índice de solvencia	3,407	0,953
Margen neto	0,036	-1,265
Rentabilidad sobre el patrimonio	0,084	0,323
Rentabilidad sobre el activo	0,043	-0,197
Nivel de endeudamiento	0,502	1,197
Apalancamiento total	2,660	-4,083
Apalancamiento a corto plazo	1,655	-2,664
Rotación activo total	1,125	1,316
Rotación cartera	8,280	8,617

Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016.*

El modelo estimado por regresión logística queda de la siguiente forma:

$$\boldsymbol{P_i = Pr(y_i = 1)} = \frac{e^{\begin{array}{c}(-14{,}741-0{,}003025Razcor_i-0{,}006691Pruacida_i+0{,}09372Razef_i \\ +0{,}49625Solv_i-0{,}01845Maneto_i-0{,}13362Roe_i-4{,}70014Roa_i+13{,}836Endeud_i \\ -0{,}040770Leveraget_i-0{,}0103810Leveragecp_i+0{,}064014rotat_i+0{,}0009050rotcar_i)\end{array}}}{1+e^{\begin{array}{c}(-14{,}741-0{,}003025Razcor_i-0{,}006691Pruacida_i+0{,}09372Razef_i \\ +0{,}49625Solv_i-0{,}01845Maneto_i-0{,}13362Roe_i-4{,}70014Roa_i+13{,}836Endeud_i \\ -0{,}040770Leveraget_i-0{,}0103810Leveragecp_i+0{,}064014rotat_i+0{,}0009050rotcar_i)\end{array}}}$$

Una vez aplicado el modelo resultante al conjunto de datos de evaluación, se presenta que el modelo clasifica 3.072 empresas de manera correcta de entre 3.105, indicando un poder predictivo general del modelo del ***98,94%,*** como se indica en la *tabla 9.* Así, el modelo clasifica de manera correcta 83,64% a las empresas frágiles y en un 99,80% para las no frágiles.

Tabla 9: Poder predictivo del modelo Logit

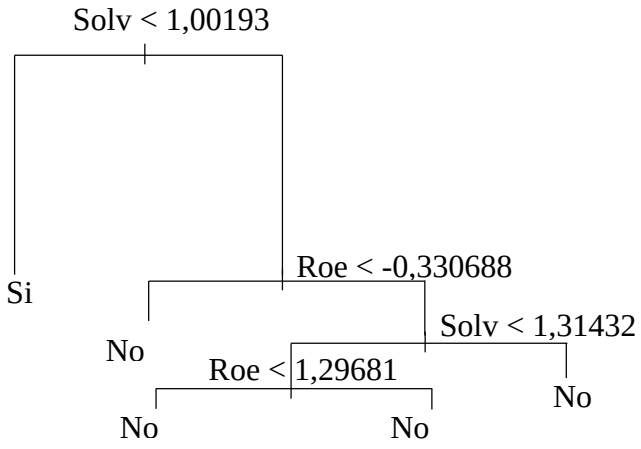
	Conjunto de evaluación	Clasificadas correctamente
No frágiles	2940	2934
Frágiles	165	138
Total	3105	3072
Poder predictivo		
Clasificación empresas no frágiles		99,80%
Clasificación empresas frágiles		83,64%
Clasificación total		98,94%

Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Árboles de clasificación

La construcción del árbol se lleva a cabo mediante un algoritmo de partición binaria, en función del índice de Gini (*ecuación 3*). En el *gráfico 1* se aprecia la estructura del árbol resultante, donde se observa que la variable más importante para la predicción de la fragilidad financiera de acuerdo al modelo de árboles de decisión la solvencia, pues es la que se encuentra en la posición más alta del árbol final, seguida de la rentabilidad sobre el patrimonio. Las demás variables, al no considerarse dentro de la gráfica, resultaron no ser de utilidad para explicar la probabilidad o no de quiebra de las empresas colombianas.

Gráfico 1: Árbol de decisión de dos nodos



Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Ahora, si bien el árbol resultante no es extenso, consideremos la poda del árbol esperando mejorar resultados y la interpretación si es posible. Por tanto, se ejecuta la validación cruzada para determinar el nivel óptimo de complejidad del árbol y elegir el tamaño de árbol óptimo (*ecuación 4 y 5*). La *tabla 10* se relaciona la información hallada de ese proceso: el número de nodos terminales de cada árbol considerado (tamaño), la tasa de error de cada árbol y el valor del parámetro costo-complejidad utilizado.

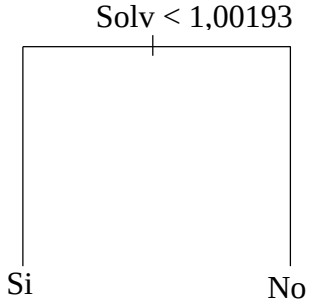
Tabla 10. Parámetro de complejidad del árbol y tamaño óptimo del árbol

Tamaño	Tasa de error	Parámetro de complejidad
5	198	
2	198	463
1	661	

Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Se halla que tanto el árbol con cinco nodos (árbol inicial) y el árbol de dos nodos resultan con el más bajo error de clasificación mediante validación cruzada, con 198 errores de validación. Estos dos presentan exactamente el mismo poder predictivo, por lo tanto, cualquiera de estos dos árboles puede ser considerado; de hecho, si se comparan, el de dos nodos resulta ser la síntesis del árbol de cinco. Debido a que luego de la poda no hubo mejora en exactitud del modelo, pero si en la interpretación, el árbol final elegido es el de dos nodos el cual arroja un modelo como se muestra a continuación en la en el *gráfico 2* e interpretado como se observa en la *tabla 11*.

Gráfico 2:Árbol de decisión de dos nodos



Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Tabla 11: Interpretación del modelo de árbol de decisión de dos nodos

Solvencia	Predicción
< 1,00193	Si
>= 1,00193	No

Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

En conclusión, para el modelo de árboles, un índice de solvencia menor a 1,00193 indica situación de fragilidad financiera en las empresas colombianas, mientras un resultado mayor o igual a 1,00193 simboliza una compañía no frágil. Con este enfoque, y aplicando el modelo a los datos de evaluación para valorar el desempeño del modelo obtenido, éste presenta una clasificación correcta de empresas no frágiles en un 99,90% y de las empresas frágiles de 84,24%, conduciendo a un poder predictivo general del modelo del ***99,07%,*** como se indica en la *tabla 12.*

Tabla 12: Poder predictivo del modelo de árboles de decisión

	Conjunto de evaluación	Clasificadas correctamente
No frágiles	2940	2937
Frágiles	165	139
Total	3105	3076

Poder predictivo	
Clasificación empresas no frágiles	99,90%
Clasificación empresas frágiles	84,24%
Clasificación total	99,07%

Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Boosting

De acuerdo a las estadísticas obtenidas con la metodología de boosting en R y utilizando como criterio la influencia relativa que tiene cada una de las variables, relacionadas en la *tabla 13,* la cual hace referencia a la importancia de cada una de las variables, se observa que el índice de

solvencia, el nivel de endeudamiento y, en menor medida, la rentabilidad sobre el patrimonio, la rentabilidad sobre el activo y el apalancamiento total, son las variables más importantes para la evaluación de la fragilidad financiera de las empresas colombianas.

Tabla 13:Influencia relativas de las variables explicativas en boosting

Variable	Influencia relativa
Índice de solvencia	48,9036
Nivel de endeudamiento	48,5455
Rentabilidad sobre el patrimonio	1,8536
Rentabilidad sobre el activo	0,6674
Apalancamiento total	0,0299
Razón corriente	0
Prueba ácida	0
Razón de efectivo	0
Margen neto	0
Apalancamiento a corto plazo	0
Rotación activo total	0
Rotación cartera	0

Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Debido a que la el boosting tiene un comportamiento de caja negra, los resultados obtenidos no se prestan para una interpretación de probabilidad de quiebra de las empresas; es decir, a diferencia de regresión logística y árboles de decisión, boosting no genera un modelo final, sino un algoritmo, en este caso compuesto por unos parámetros shrinkage (*λ)= 0,001, depth (número de divisiones) =2 y una combinación de 2000 árboles.* Con esos criterios se aplican las *ecuaciones 6, 7, 8 y 9*.

Los parámetros del algoritmo fueron seleccionados mediante el paquete estadístico R, de tal forma que para maximizaran el poder predictivo del modelo. El algoritmo se aplicó y dio como resultado un poder predictivo general del 99%, de donde la clasificación correcta de las empresas no frágiles fue del 99,80% y de las frágiles, un 84,85% (*ver tabla 14*).

Tabla 14: Poder predictivo del modelo estimado con boosting

	Conjunto de evaluación	Clasificadas correctamente
No frágiles	2940	2934
Frágiles	165	140
Total	3105	3074

Poder predictivo	
Clasificación empresas no frágiles	99,80%
Clasificación empresas frágiles	84,85%
Clasificación total	99,00%

Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Comparación entre los modelos

Para evaluar y comparar el desempeño de los tres modelos se utiliza inicialmente la métrica expuesta en la *tabla 15*. En esta se observa en general un desempeño muy bueno en cada uno de los modelos. A simple vista, el que menor rendimiento, siendo todavía alto, es regresión logística, dado que presenta tasa de aciertos, especificidad y sensibilidad más bajas y tasa de errores, tasa de falsos unos y tasa de falsos ceros mayores. En cuanto a los dos modelos restantes, es difícil determinar cuál es mejor debido a que si bien, árboles de decisión presenta una tasa de aciertos, y una especificidad mayor un poco más elevada que boosting, este último presenta una sensibilidad ligeramente mayor.

Tabla 15. Comparación de modelos

Índices	Logit	Árbol de clasificación	Boosting
Tasa de aciertos	98,94%	99,07%	99,00%
Tasa de errores	1,06%	0,93%	1,00%
Especificidad	99,80%	99,90%	99,80%
Sensibilidad	83,64%	84,24%	84,85%
Tasa de falsos ceros	0,2041%	0,1020%	0,2041%
Tasa de falsos uno	16,36%	15,76%	15,15%

Fuente: *elaboración propia en base a la información reportada por la superintendencia de Sociedades de Colombia a diciembre 31 de 2016, usando el paquete estadístico R.*

Al observar ambigüedades para elegir entre arboles de clasificación y boosting, el rendimiento de estos modelos para predecir la bancarrota se prueba finalmente a través del área bajo la curva ROC, sobre los datos de prueba de cada uno de ellos.

La curva ROC es una representación gráfica de la relación existente entre la distribución de la fracción de verdaderos positivos y de falsos positivos. La fracción de los verdaderos positivos se conoce como sensibilidad, e indica la probabilidad de clasificar de manera correcta a una empresa en quiebra. La facción de los falsos positivos es conocido como 1 menos la especificidad, probabilidad de clasificar de manera correcta a una empresa que no está en quiebra (Brîndescu–Olariu, 2016)

El área bajo la curva la curva ROC, conocida como AUC, es una de las formas más viables para medir el rendimiento de clasificación y para comparar modelos, pues representa la precisión general de la clasificación representa el porcentaje de empresas correctamente clasificadas. Ésta puede tomar valores entre 0 y 1. Un AUC de 0,5 corresponde a una precisión mala, mientras un AUC de 1 corresponde a una precisión perfecta(Brîndescu–Olariu, 2016). De acuerdo con el autor, la evaluación de precisión se evalúa según la siguiente escala:

Tabla 16. Escala de calificación AUC

Escala	Calificación
0,5 – 0,6	Malo
0,6-0,7	Pobre
0,7-0,8	Suficiente
0,8-0,9	Bueno
0,9 - 1	Excelente

Fuente: (Brîndescu–Olariu, 2016, p. 4)

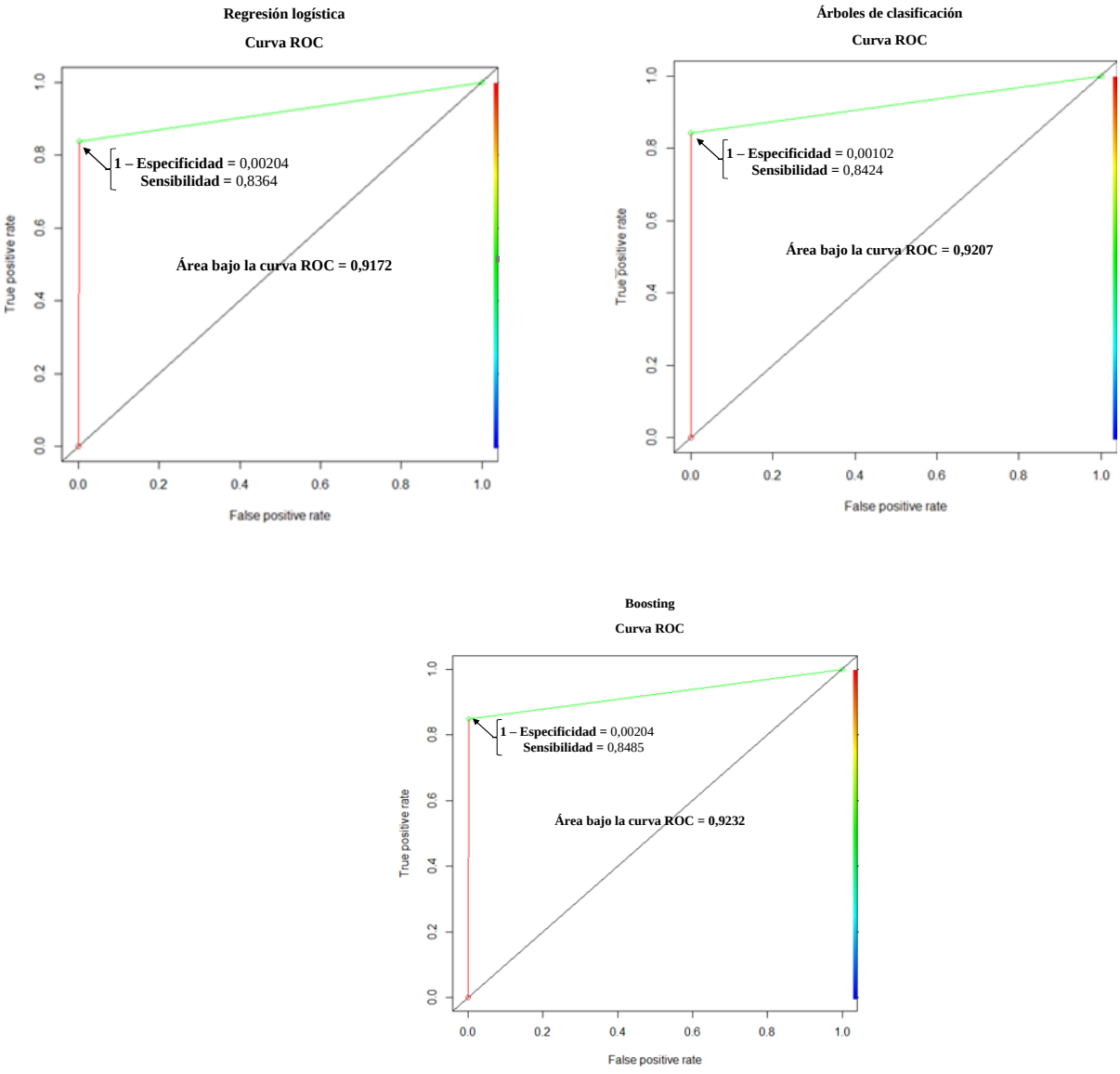
Las áreas bajo la curva ROC para uno de los modelos muestra que en general los tres pueden ser utilizados como una herramienta para la evaluación de riesgo de quiebra. Esta conclusión está sustentada en una clasificación correcta de las empresas de un 91,72% en el modelo de regresión logística, 92,07% mediante árboles de clasificación y 92,32% con boosting. Como se observa, boosting se sitúa en primer lugar en cuanto a modelo con mejor desempeño; sin embargo, las diferencias son pequeñas en porcentaje con respecto a los dos modelos restantes, especialmente con árboles de decisión, cuya diferencia es ínfima.

Tabla 17. Áreas bajo la curva ROC (AUC)

	Logit	Árbol de clasificación	Boosting
Área bajo la curva	91,72%	92,07%	92,32%

Fuente: Elaboración propia

Gráfico 1. Curvas ROC para modelos de predicción



Fuente: Elaboración propia con base en el paquete estadístico R

CONCLUSIONES

En los últimos años hubo un creciente interés en la teoría y aplicación en cuanto a sucesos raros se refiere. Dentro de esta categoría, se destacan los esfuerzos que se han realizado durante las últimas décadas en el estudio de la *bancarrota*, debido principalmente a las consecuencias en la sociedad y a la incertidumbre asociada a esta situación, máxime cuando el entorno actual es en demasía cambiante. Esto sumado a que pese a la existencia de múltiples estudios que desarrollan este tipo de modelos, su aplicación es escasa para el caso de Colombia, lo que hizo que el objetivo de este trabajo fue la estimación de tres modelos de predicción de la quiebra empresarial en base a tres metodologías diferentes: regresión logística, árboles de decisión y boosting.

El estudio arroja que las variables que mejor describen el nivel de fragilidad financiera de las empresas colombianas con el modelo logístico son el nivel de solvencia, nivel de endeudamiento, el apalancamiento total y rentabilidad sobre el activo (ROA). Los resultados obtenidos demuestran que las empresas tienen mayor propensión a la quiebra cuando poseen bajos niveles de rentabilidad en los activos, al igual que bajo apalancamiento total. También, altos resultados en el nivel de solvencia y en el nivel de endeudamiento contribuyen al incremento de la fragilidad.

Aunque el indicador de rentabilidad sobre el patrimonio (ROE) en un inicio se configuró como una variable marginalmente significativa en el modelo, analizado la significancia de los ratios en un nuevo escenario donde se deja de considerar al ROE como una variable explicativa, se obtiene una mejora en cuanto a los resultados de predicción (predice de manera correcta una empresa más en comparación del modelo inicial), y las variables estadísticamente significativas siguen siendo exactamente las mismas. Por lo tanto, se afirma la rentabilidad sobre el patrimonio no se considera dentro de los ratios más útiles para explicar el nivel de fragilidad financiera de las empresas colombianas.

En cuanto al modelo estimado mediante árboles de decisión, la variable más importante para la predicción de la fragilidad financiera la solvencia, pues es la que se encuentra en la posición más alta del árbol final. Las demás variables, al no considerarse dentro de la gráfica, resultaron no ser de utilidad para explicar la probabilidad o no de quiebra de las empresas colombianas. Conjuntamente, se dedujo mediante esta técnica que un índice de solvencia menor a 1,00193 indica

situación de fragilidad financiera en las empresas colombianas, mientras un resultado mayor o igual a 1,00193 simboliza una compañía no frágil.

Por último, la técnica de boosting permitió determinar al nivel de solvencia, nivel de endeudamiento, rentabilidad sobre el patrimonio, la rentabilidad sobre el activo y apalancamiento total, como los índices que mejor explican la fragilidad financiera.

En conclusión, estimando un modelo Logit, un modelo de árboles de decisión y la utilización de boosting, se identificó como la variable más significativa para medir la probabilidad de quiebra empresarial en Colombia, al índice de solvencia, indicador que resultó ser común en los tres modelos estimados, seguido del nivel de endeudamiento que fue notablemente significativo en dos de ellos. La primera, establece la facilidad o dificultad de la empresa de cumplir con sus obligaciones (tanto operativas como financieras) tanto de corto como de largo plazo si se le exigiese el pago inmediato de éstas, y la segunda, el nivel de participación tienen los acreedores de entro de la empresa.

Al evaluar el poder predictivo de cada uno de los modelos mediante la tasa de evaluación, la sensibilidad y la especificidad, se encuentra que el modelo que presenta menor poder predictivo, siendo todavía muy alto, es regresión logística. Para terminar de confirmar, y al observar ambigüedades para elegir el modelo que mejor rendimiento presenta para predecir la bancarrota, entre arboles de clasificación y boosting, se utiliza finalmente la métrica basada en el área bajo la curva ROC, sobre el conjunto evaluación.

Las áreas bajo la curva ROC para uno de los modelos muestra que en general los tres pueden ser utilizados como una herramienta para la evaluación de riesgo de quiebra. Esta conclusión está sustentada en una clasificación correcta de las empresas de un 91,72% en el modelo de regresión logística, 92,07% mediante árboles de clasificación y 92,32% con boosting. Como se observa, boosting se sitúa en primer lugar en cuanto a modelo con mejor desempeño; sin embargo, las diferencias son pequeñas en porcentaje son pequeñas con respecto a los dos modelos restantes, especialmente con árboles de decisión, cuya diferencia es ínfima.

A pesar de la aparente similitud en los resultados arrojados de cada modelo, se puede argumentar que una muestra de 15.524 datos, como en este caso, no es suficiente para evidenciar un delta significativo entre los resultados de los modelos; por lo que muestras mucho más grandes, la brecha puede hacerse más evidente. Con ello se demuestra que, a pesar de que las primeras dos

son buenas herramientas predictivas, boosting se consolida como una metodología con mejor poder predictivo, aunque computacionalmente más exigente.

Si bien a nivel teórico árboles de decisión es más potente frente a regresión logística dado tiene un carácter no paramétrico que le otorga superioridad en cuanto a la facilidad de adaptación a los datos, en este caso se observan resultados muy parecidos. Además, a pesar de que existen estudios que indican que la técnica de regresión logística no es buena herramienta de predicción cuando la variable dependiente es catalogada como un evento raro, como es el caso de la quiebra, para el conjunto de datos de la muestra seleccionada para el trabajo, el comportamiento dista de la premisa. Los resultados arrojan una tasa de clasificación correcta del 98,94% y un área bajo la curva de un 91,72%, que si bien están por debajo del rendimiento de árboles de clasificación y boosting, no deja de pertenecer a la categoría "excelente precisión" de acuerdo a la escala de evaluación descrita por Brîndescu–Olariu (2016)

Por ello, no puede afirmarse ligeramente que existan mejores y peores metodologías, sino que cuando se construyen estos tipo de modelos, la elección de factores como las variable dependiente (definición de quiebra), las variables explicativas y la cantidad de la muestra, es subjetiva, lo cual interviene de manera crucial en los resultados, pues como menciona Tascón & Castaño (2012), que funcione mejor uno u otro modelo depende del conjunto de datos sobre el que se aplique.

Además, dado que la metodología Logit es muy sensible a multicolinealidad, a la existencia de valores extremos y a la falta de datos desaparecidos, el buen desempeño del modelo Logit en este estudio también se le adjudica al tratamiento que se hizo de los datos, tratando cada uno de esos aspectos.

Si bien la predicción de los modelos resulto ser muy buena, este estudio se enfocó solo en 12 ratios financieros como base para estimar los modelos. Por ello, como recomendación a tener en cuenta en la realización de futuros estudios de este tipo se hace alusión a lo expuesto por Tascón & Castaño (2012) en cuanto a uno de los notables limitantes en la aplicación de las metodologías de predicción de fracaso empresarial, el cual radica en la ausencia de factores externos a las organizaciones como variables explicativas en la mayoría de los estudios, de modo que en situaciones de cambio estos modelos no detectan la fragilidad financiera sino hasta que se ven reflejados en la información contable, conllevando a tardar en detectar problemas. Asimismo, sería

conveniente considerar periodos adicionales a evaluar para comprobar la capacidad explicativa de los indicadores financieros y el poder predictivo de los modelos en tiempos distintos.

Es importante mencionar que la intención del estudio no es sustituir la experiencia ni la capacidad de discernimiento de los encargados de evaluar la situación financiera dentro de las organizaciones, pero si complementar el proceso de detección de quiebra. Además, no existe regla general para evitarla; sin embargo, estos modelos pueden ser muy beneficiosos para ayudar a detectar y prevenir una situación financiera difícil con antelación pues dados los resultados de la aplicación del modelo a cada caso, se puede impulsar un proceso interno de análisis y evaluación de la situación actual y posible situación futura, que permitan la detección temprana de la quiebra y la toma de medidas correctivas a tiempo.

REFERENCIAS

Adnan, M., & Dar, H. A. (2006). Predicting corporate bankruptcy: where we stand? *Corporate Governance: The International Journal of Business in Society*, *6*(1), 18–33. https://doi.org/10.1108/14720700610649436

Alaminos, D., Del Castillo, A., & Fernández, M. Á. (2016). A Global Model for Bankruptcy A Global Model for Bankruptcy Prediction. *Plos One*, *11*(11), 1–18. https://doi.org/10.1371/journal.pone.0166693

Beaver, W., McNichols, M., & Rhie, J. W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies*, *10*(1), 93–122. https://doi.org/10.1007/s11142-004-6341-9

Bourel, M. (2012). Métodos de agregación de modelos y aplicaciones. *Memoria de Trabajos de Difusión Científica Y Técnica*, *10*, 19–32.

Bouza, C. N., & Santiago, A. (2012). La minería de datos: árboles de decisión y su aplicación en estudios médicos. *Modelación Matématica de Fenómennos Del Medio Ambiente Y La Salud*, *2*, 64–78.

Brîndescu–Olariu, D. (2016). Solvency ratio as a tool for bankruptcy prediction. *Ecoforum Journal*, *5*(2), 278–281. Retrieved from

http://ecoforumjournal.ro/index.php/eco/article/view/418/265

Contreras, J. G. (2016). *Análisis de quiebra empresarial: modelo de ecuaciones de estimación generalizadas sobre datos panel*. Universidad Complutense de Madrid.

Cruz, J. S., Villareal, J., & Rosillo, J. S. (2001). *Finanzas corporativas - Valoración, política de financiamiento y riesgo*. (International Thomson Editores., Ed.). Bogotá, colombia: International Thomson Editores.

Diaz, Z. (2000). *Predicción de crisis empresariales en seguros de vida, mediante árboles de decisión y reglas de clasificación*. Madrid, españa: Editorial Complutense.

Diaz, Z., Fernandez, J., & Segovia, M. J. (2004). Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras. Madrid, España.

Elam, R. (1975). The Effect of Lease Data on the Predictive Ability of Financial Ratios. *The Accounting Review*, *50*(1), 25–43. Retrieved from http://libaccess.mcmaster.ca.libaccess.lib.mcmaster.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4501571&site=ehost-live&scope=site

Gestel, T., Baesens, B., Suykens, J., Poel, D., Baestaens, D., & Willekens, M. (2006). Bayesian kernel based classification for financial distress detection. *European Journal of Operational Research*, *172*(3), 979–1003.

Goleţ, I. (2014). Symmetric and Asymmetric Binary Choice Models for Corporate Bankruptcy. *Procedia - Social and Behavioral Sciences*, *124*, 282–291. https://doi.org/10.1016/j.sbspro.2014.02.487

Gujarati, D. N., & Porter, D. C. (2013). *Econometría*. *Mc Graw Hill* (Vol. 53). México DF, México: Mc Graw Hill. https://doi.org/10.1017/CBO9781107415324.004

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to Statistical Learning-whith applications in R*. (G. Casella, S. Fienberg, & I. Olkin, Eds.) (Vol. 7). New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Li, Z., Crook, J., & Andreeva, G. (2014). Chinese companies distress prediction: An application of data envelopment analysis. *Journal of the Operational Research Society*, *65*(3), 466–479. https://doi.org/10.1057/jors.2013.67

Martínez, O. (2003). Particular Aspects of Financial Stability. Determinants of Corporate Fragility

in Colombia. *Borradores de Economía*. Bogotá, colombia: Banco de la República. Retrieved from http://www.banrep.gov.co/es/borrador-259

Mora, A. E. (1994). Limitaciones metodológicas de los trabajos empíricos sobre la predicción del fracaso empresarial. *Revista Española de Financiación Y Contabilidad*, *XXIV*(80), 709–732.

Mulyawan, S. (2015). The Benefits of Financial Ratios' as the Indocators of Future Bankruptcy on the Economic Crisis. *International Journal of Nusantara Islam*, *3*(1), 21. Retrieved from https://doi.org/10.15575/ijni.v3i1.153

Ortiz, H. (1998). *Análisis financiero aplicado - con ajustes por inflación*. (Universidad Externado de Colombia, Ed.) (10th ed.). Bogotá, colombia: Universidad Externado de Colombia.

Perez, J., Gonzalez, K., & Lopera, M. (2013). Modelos De Prediccion De La Fragilidad Empresarial: Aplicacion Al Caso Colombiano Para El Año 2011. *Perfil de Coyuntura Económica*, (22), 205–228. Retrieved from http://www.redalyc.org/pdf/861/86131758010.pdf

Ringeling, E. A. (2004). *Análisis comparativo de modelos de predicción de quiebra y la probabilidad de bancarrota*. Universidad de Chile. Retrieved from http://www.tesis.uchile.cl/tesis/uchile/2004/ringeling_e/sources/ringeling_e.pdf

Rosillo, J. (2002). Modelo de predicción de quiebras de las empresas colombianas. *Innovar, Revista de Ciencias Administrativas Y Sociales*, *19*(19), 109–124. Retrieved from http://www.bdigital.unal.edu.co/26382/1/23959-83815-1-PB.pdf

Seoane, J., Carmona, C., Tarjuelo, R., & Planillo, A. (2014). Análisis bioestadístico con modelos de regresión en R. Retrieved from http://www.uam.es/personal_pdi/ciencias/jspinill/CFCUAM2014/RF_BRT-CFCUAM2014.html

Serna, C. (2009). *Comparación de Árboles de Regresión y Clasificación y regresión logística (tesis de maestría)*. Univeridad Nacional de Colombia. Retrieved from http://www.bdigital.unal.edu.co/671/1/42694070_2009.pdf

Supersociedades. (2012). Causas de la Insolvencia Empresarial. *Revista Supersociedades*, (4), 27–31.

Tascón, M. T., & Castaño, F. J. (2012). Variables y modelos para la identificación y predicción del fracaso empresarial: Revisión de la investigación empírica reciente. *Revista de Contabilidad*, *15*(1), 7–58. https://doi.org/10.1016/S1138-4891(12)70037-7

Wu, Y., Gaunt, C., & Gray, S. (2010). A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting and Economics*, *6*(1), 34–45. Retrieved from https://doi.org/10.1016/j.jcae.2010.04.002