# Modelo de alerta de quiebra probable a dos años para empresas colombianas utilizando algoritmos de *machine learning*

## Probable Bankruptcy Alert Model at two Years for Colombian Companies Using Machine Learning Algorithms

Julián Andrés Villamizar Peñaranda*

* Magíster en Finanzas, Universidad Externado de Colombia. Head of Supply Chain Efficacy, Bogotá (Colombia). [julian.villamizar@gmail.com]
Artículo recibido: 23 de julio de 2024
Aceptado: 20 de septiembre de 2024

Para citar este artículo:
Villamizar Peñaranda, J. A. (2024). Modelo de alerta de quiebra probable a dos años para empresas colombianas utilizando algoritmos de *machine learning*. *Odeon*, *27*, 209-254.
DOI: https://doi.org/10.18601/17941113.n27.07

## Resumen

En este estudio se desarrolla un modelo predictivo de quiebra empresarial para empresas colombianas, basado en algoritmos de *machine learning*. A diferencia de modelos tradicionales enfocados en la predicción de quiebra a corto plazo, este trabajo propone un modelo que proyecta la quiebra a dos años. El modelo fue construido usando datos financieros suministrados por la Superintendencia de Sociedades, los cuales fueron depurados para adaptarse a la realidad del entorno empresarial colombiano, excluyendo variables como el EBITDA y la capitalización bursátil por falta de información. Se usaron como base teórica los modelos de Altman y las mejoras más recientes al Z-Score, incorporando indicadores adicionales como el Total Activo/Total Pasivo y el Pasivo a Corto Plazo/Total Pasivo. El algoritmo de Random Forest mostró un área bajo la curva (AUC) del 92,89% para predicciones a dos años, lo que confirma su utilidad como una herramienta de alerta temprana para la gerencia financiera.



**Palabras clave:** predicción de quiebra; *machine learning*; Random Forest; Altman Z-Score; empresas colombianas.



## Abstract

This study develops a predictive bankruptcy model for Colombian companies using Machine Learning algorithms. Unlike traditional models focused on short-term bankruptcy prediction, this work proposes a model that projects bankruptcy two years ahead. The model was built using financial data provided by the Superintendency of Companies, which was refined to suit the Colombian business environment, excluding variables like EBITDA and market capitalization due to lack of information. The theoretical basis includes Altman's models and recent improvements to the Z-Score, incorporating additional indicators such as Total Assets/Total Liabilities and Short-term Liabilities/Total Liabilities. The Random Forest algorithm demonstrated an area under the curve (AUC) of 92.89 % for two-year predictions, confirming its utility as an early warning tool for financial management.



**Keywords:** Bankruptcy prediction; Machine Learning; Random Forest; Altman Z-Score; Colombian companies.

**JEL classification:** C45, G33, M41

## Introducción

La preocupación por el futuro es una constante en el ser humano; el temor a lo desconocido y la posibilidad de pérdida llevan a un comportamiento de aversión al riesgo, el cual ha caracterizado las decisiones de inversión a lo largo de la historia. Este comportamiento se refleja incluso en la teoría de portafolio de Markowitz, donde se asume que los agentes son aversos al riesgo y buscan la mayor rentabilidad al menor riesgo posible. Con base en este principio, los inversionistas y académicos han trabajado en el desarrollo de modelos para anticipar la quiebra empresarial y evitar invertir en compañías con alto riesgo, asegurando así retornos futuros positivos.

Históricamente, estos modelos se han basado en la información contable de las empresas, evolucionando desde decisiones sustentadas en indicadores financieros hasta el uso de modelos matemáticos más complejos y, más recientemente, la implementación de algoritmos de *machine l*earning *(ML)* e inteligencia artificial (IA). En los últimos años, estos desarrollos han alcanzado un nuevo nivel gracias al aumento en el poder computacional y la capacidad de procesamiento de datos, permitiendo que los modelos de ML e IA se conviertan en los principales protagonistas en la predicción de quiebra. No obstante, este avance no habría sido posible sin la disminución en los costos de almacenamiento y la disponibilidad de datos históricos proporcionados por empresas privadas o entidades gubernamentales.

Dado que se dispone de estos mecanismos para desarrollar modelos robustos sin necesidad de grandes inversiones en la obtención de datos, el presente trabajo propone adaptar los más recientes desarrollos en modelos de predicción de quiebra a la información disponible en el contexto colombiano. Sin embargo, este proceso enfrenta retos significativos. Entre ellos, la calidad y disponibilidad de los datos financieros reportados por la Superintendencia de Sociedades de Colombia, los cambios en los estándares contables, y la dificultad de contar con indicadores financieros confiables y detallados. A pesar de estas limitaciones, se espera que, con una adecuada depuración de los datos se obtenga un modelo con una precisión comparable con aquellos aplicados en economías más desarrolladas.

Los modelos de predicción de quiebra han recorrido un largo camino y, aunque se ha avanzado considerablemente, aún no existe uno que logre predecir la mayoría de las quiebras de manera efectiva. Esto subraya la necesidad de seguir mejorando los modelos, especialmente en economías emergentes como

la de Colombia. Estas economías enfrentan retos específicos en cuanto a la disponibilidad y calidad de la información, y los modelos que funcionan bien en países desarrollados suelen incluir variables del mercado de capitales, las cuales actúan como indicadores líderes basados en expectativas. Sin embargo, en Colombia, el mercado accionario es poco desarrollado, con un número reducido de empresas cotizando en bolsa, lo que obliga a prescindir de estas variables y a buscar alternativas.

Este trabajo toma como base el reconocido modelo Z-Score de Edward Altman, el cual ha mostrado resultados sólidos en economías avanzadas. A partir de este punto de partida, se exploran otras técnicas matemáticas y las más recientes aplicaciones de algoritmos de *machine learning* para crear una solución adaptada a los retos específicos de la información disponible en Colombia. En resumen, el objetivo es desarrollar un algoritmo de ML con alta capacidad predictiva que se ajuste a las limitaciones de disponibilidad y calidad de la información financiera en el entorno colombiano.

## 1. Antecedentes teóricos

La predicción de quiebra ha sido objeto de estudio desde hace más de 50 años, cuando William H. Beaver publicó su modelo de análisis univariante en su trabajo "Financial ratios as predictors of failure" (1966). Desde entonces, numerosos investigadores han tratado de mejorar la capacidad predictiva de estos modelos, desarrollando técnicas cada vez más complejas. Sin embargo, es fundamental comenzar por definir el concepto de "quiebra" que se utilizará en este estudio. A lo largo de la historia, las definiciones han variado considerablemente, como se puede observar en la tabla 1 (Romero Espinosa, 2013), que recoge un resumen de las definiciones de quiebra según diferentes autores. Desde la incapacidad de atender deudas hasta la insolvencia técnica y la quiebra legal, cada definición ofrece una perspectiva particular del concepto. Para este trabajo, se utilizará la definición legal de quiebra según la Superintendencia de Sociedades de Colombia, que incluye la admisión a procesos de reorganización y la liquidación judicial conforme a la Ley 1116 de 2006. Esta definición es la más adecuada debido a las limitaciones para identificar negociaciones privadas entre acreedores y deudores, y está alineada con el objetivo del modelo de alerta temprana, que busca prever la quiebra con suficiente antelación para que las empresas tomen medidas correctivas (Romero Espinosa, 2013).

Tabla 1. Definiciones históricas de quiebra

| Autor | Término | Definición |
|---|---|---|
| Beaver, 1966 | Fracaso | Dificultad para atender deudas (Obligaciones financieras) |
| Altman, 1968 | Quiebra | Catalogadas legalmente en quiebra |
| Ohlson, 1980 | Quiebra | Legalmente en quiebra |
| Altman, 1981 | Quiebra | Insolvencia técnica-falta de liquidez |
| Zmijewski, 1984 | Quiebra | Quiebra legal |
| Zavgren, 1985 | Quiebra | Quiebra legal, suspensión de pagos |
| Lo, 1986 | Quiebra | Legalmente en quiebra |
| Altman, 1988 | Quiebra | No pueda hacer frente a sus obligaciones con los acreedores |
| Theodossiou, 1993 | Quiebra | Insolvencia, legalmente en quiebra |
| Correa, Acosta, González, 2003 | Quiebra | Patrimonio negativo o quiebra técnica |
| Rubio Misas, 2008 | Quiebra | Patrimonio negativo o quiebra técnica |

Fuente: Romero Espinosa (2013).

Los modelos de predicción de quiebra han evolucionado significativamente desde los primeros análisis univariantes. La tabla 2 proporciona una vista general de la evolución histórica de estos modelos (Romero Espinosa, 2013), comenzando con el análisis univariante de Beaver en 1966 y siguiendo con los avances de Edward Altman (1968) con su análisis discriminante múltiple, hasta llegar al uso de algoritmos de *machine learning* (*ML*) en las últimas décadas. Los avances en el poder computacional y en el procesamiento de datos han permitido el desarrollo de modelos más sofisticados, como los algoritmos de *ML*, que han mostrado una mayor precisión en la predicción de quiebras (Trappenberg, 2020). A lo largo de los años, se ha demostrado que las variables seleccionadas para alimentar estos modelos tienen un impacto crucial en su capacidad predictiva, y cualquier cambio en las variables puede alterar significativamente los resultados (Romero Espinosa, 2013).

Tabla 2. Historia de los modelos usados para predecir la quiebra empresarial

| Año | Autor | Metodología utilizada |
|---|---|---|
| 1966 | Beaver | Análisis univariantes |
| 1968 | Altman | Análisis discriminante múltiple |
| 1972 | Deakin | Análisis discriminante múltiple |
| 1980 | Ohlson | Análisis de regresión logística |
| 1984 | Marais, Patell y Wolfson | Modelos de partición condicional |
| 1985 | Zavgren | Análisis de probabilidad condicional, Análisis de regresión logística |
| 1987 | Goudie | Análisis discriminante múltiple sectorial |
| 1988 | Dutta y Shekkar | Inteligencia artificial |
| 1990 | Odom y Sharda | Inteligencia Artificial: Redes neuronales |
| 1991 | Platt y Platt | Análisis de regresión logística |
| 1993 | Theodossiou | Análisis discriminante *vs*. Suma acumulante |
| 1998 | Ferrando y Blanco | Análisis discriminante y Logit |
| 2006 | Calvo-Flores, García y Madrid | Análisis de regresión logística |
| 2009 | Xu y chang | Análisis discriminante múltiple, Análisis de regresión logística |

Fuente: Romero Espinosa, (2013).

Uno de los modelos más importantes en la historia de la predicción de quiebras es el Altman Z-Score, desarrollado en 1968, que se basa en el análisis discriminante múltiple y ha sido ampliamente utilizado en estudios posteriores. Este modelo emplea cinco variables financieras clave, incluyendo la capitalización bursátil y la razón de ventas sobre activos totales, para calcular una puntuación que clasifica a las empresas en tres zonas: segura, gris o de peligro (Altman, 1968). El Z-Score ha servido como base para muchos desarrollos posteriores, incluidos los trabajos de Barboza *et al*. (2017), quienes mejoraron el modelo al incorporar nuevas variables y técnicas de *ML*, como el Random Forest. Estas mejoras, que incluyeron indicadores financieros adicionales como el margen operativo y el crecimiento de las ventas, aumentaron la capacidad predictiva del

modelo, particularmente en entornos con grandes volúmenes de datos (Romero Espinosa, 2013).

Los modelos de *ML* han demostrado ser particularmente efectivos para la predicción de quiebras debido a su capacidad para manejar conjuntos de datos grandes y complejos (Barboza *et al*., 2017). Técnicas como los árboles de decisión, las redes neuronales y las máquinas de soporte vectorial (SVM) han ampliado el campo de la predicción de quiebras, brindando herramientas más precisas que los métodos tradicionales (Trappenberg, 2020). El trabajo de Barboza *et al*. es especialmente relevante, ya que en su estudio compararon múltiples algoritmos, incluyendo SVM, Redes Neuronales, Logit, y Random Forest, obteniendo los mejores resultados con este último. El uso de Random Forest en combinación con variables adicionales permitió un aumento significativo en el área bajo la curva ROC (AUC) y una mayor precisión en la clasificación de empresas en quiebra, superando a los modelos tradicionales (Barboza *et al*., 2017).

A pesar de los avances en los métodos de predicción de quiebras, los retos siguen siendo significativos, especialmente en economías emergentes como en Colombia, donde la disponibilidad de datos financieros detallados es limitada. Esto plantea un desafío para la adaptación de modelos como el Z-Score, que requieren variables como la capitalización bursátil, la cual no está disponible en la mayoría de las empresas colombianas. Sin embargo, estudios recientes han demostrado que con la correcta selección de variables alternativas, es posible desarrollar modelos adaptados a estos contextos. En este sentido, la aplicación de técnicas de *machine learning*, y en particular el uso del Random Forest, se presenta como la opción más prometedora para la predicción de quiebras en empresas colombianas (Barboza *et al*., 2017).

En resumen, los desarrollos en la predicción de quiebras han recorrido un largo camino desde los modelos univariantes hasta los algoritmos de *machine learning* actuales. A través del análisis de técnicas como el Altman Z-Score y las mejoras introducidas por investigadores como Barboza, Kimura y Altman, se ha demostrado que los modelos basados en *ML*, y especialmente el Random Forest, ofrecen una capacidad predictiva superior, ajustándose mejor a las realidades de las economías emergentes. Sin embargo, como se ha mencionado, la efectividad del modelo dependerá en gran medida de la calidad y disponibilidad de los datos utilizados para entrenar los algoritmos, y en este sentido, la

adaptación de los modelos tradicionales a contextos como el colombiano sigue siendo un desafío (Romero Espinosa, 2013; Barboza *et al*., 2017).

## 2. Metodología aplicada

Tras la revisión de los modelos desarrollados anteriormente y el análisis de sus resultados, se determinó que el Altman Z-Score sería un punto de partida sólido, y que el Random Forest es el algoritmo más adecuado para la predicción de quiebra en el contexto colombiano. El Random Forest ha demostrado ser robusto en cuanto a su capacidad para manejar variables continuas y evitar el sobreajuste al utilizar múltiples árboles de decisión con diferentes rutas y configuraciones, lo que aumenta la precisión de las predicciones (Barboza *et al*., 2017).

El estudio de Barboza *et al*. (2017) demostró el poder del Random Forest para predecir la quiebra a un año, por lo que este algoritmo se aplicará a tres conjuntos de datos de empresas colombianas. El primer conjunto de datos se usará para entrenar y evaluar el modelo utilizando las variables originales del Altman Z-Score. El segundo conjunto incluirá las variables del Z-Score más las adicionales propuestas por Barboza *et al*. (2017), mientras que el tercer conjunto agregará tres nuevas variables específicas para empresas colombianas: Total Activo/Total Pasivo, Pasivo Corto Plazo/Total Pasivo, y Razón Corriente. Este último conjunto será la propuesta final para la predicción de quiebras en el contexto colombiano.

Debido a la falta de datos de capitalización bursátil y número de empleados, estas variables se eliminaron del análisis. También se ajustaron los cálculos que dependían de la utilidad operativa y las utilidades retenidas, sustituyéndolos por la utilidad neta debido a inconsistencias en los datos financieros reportados por la Superintendencia de Sociedades (Romero Espinosa, 2013).

El modelo original del Z-Score de Altman (1968) incluye cinco variables financieras clave, pero debido a la calidad y disponibilidad de datos en Colombia, se hicieron varios ajustes. Se eliminó la variable de Capitalización Bursátil/Total Pasivos y se reemplazó el cálculo de EBIT/Activos por Utilidad Neta/Activos Totales, haciéndolo más sensible a los gastos financieros y la tasa de impuestos. Las cuatro variables restantes se mantienen como parte del primer conjunto de datos:

- Capital de Trabajo/Activos Totales
- Utilidades Retenidas/Activos Totales
- Utilidad Neta/Activos Totales
- Ventas Totales/Activos Totales

### 2.1 Ajustes del modelo de Barboza *et al.*, para Colombia

En el documento *Machine Learning models and Bankruptcy Prediction* (Barboza *et al*., 2017), se incluyeron seis variables adicionales a las del Z-Score. Sin embargo, dos de estas (el cambio en la relación precio/valor en los libros y el crecimiento en empleados) no pudieron ser utilizadas debido a la falta de información de mercado y de reportes laborales en Colombia. No obstante, el crecimiento en ventas y activos, que agregan información de tendencias, fueron conservados. Además, se reemplazó el margen operativo por el margen neto debido a la inconsistencia en los datos de utilidad operativa.

Las variables adicionales incluidas en este segundo conjunto de datos son:

- Margen Neto
- Retorno sobre el Patrimonio
- Crecimiento en Ventas
- Crecimiento de Activos

Para mejorar la capacidad predictiva de los dos primeros conjuntos de datos, se incorporaron tres variables adicionales en el tercer conjunto:

- Total Activo/Total Pasivo: indica el nivel de apalancamiento de la empresa.
- Pasivo Corto Plazo/Total Pasivo: mide la concentración de la deuda a corto plazo, que puede ser un indicador de problemas de liquidez.
- Razón Corriente: indicador ampliamente utilizado para evaluar la liquidez de una empresa.

El conjunto final de variables evaluadas es el siguiente:

- Capital de Trabajo/Activos Totales
- Utilidades Retenidas/Activos Totales

- Utilidad Neta/Activos Totales
- Ventas Totales/Activos Totales
- Margen Neto
- Retorno sobre el Patrimonio
- Crecimiento en Ventas
- Crecimiento de Activos
- Total Activo/Total Pasivo
- Pasivo Corto Plazo/Total Pasivo
- Razón Corriente

### 2.2 Metodologías de evaluación

Para evaluar el desempeño del modelo de Random Forest, se aplicarán varias métricas de clasificación y validación siguiendo las recomendaciones de Thomas P. Trappenberg (2020). Entre ellas:

- Matriz de confusión: evalúa las predicciones correctas e incorrectas para ambas categorías: quiebra y estable.
- Exactitud (*Accuracy*): proporción de predicciones correctas en general, sin distinguir entre quiebra y estabilidad.
- *Recall*: mide la proporción de quiebras correctamente predichas.
- Precisión: indica qué tan bien predice el modelo las quiebras sin sobreestimar la clase positiva.
- F1 Score: media armónica entre la precisión y el *recall*, que equilibra ambas métricas.
- Curva ROC y AUC: la curva ROC mide la sensibilidad frente a la especificidad, mientras que el AUC cuantifica el poder predictivo del modelo. Un valor más cercano a 1 indica una mayor capacidad predictiva.

Con estas herramientas, se espera evaluar la capacidad del modelo para identificar de manera temprana las empresas en riesgo de quiebra y proporcionar alertas que permitan a los administradores tomar decisiones proactivas.

## 3. Fuente de los datos

En el presente trabajo se usaron los reportes empresariales de la Superintendencia de Sociedades de empresas del grupo 1 y grupo 2. De los datos disponibles se

seleccionaron los años 1995 a 2015 para el set de entrenamiento y 2016 para el set de evaluación. Se excluyeron datos de 2017 en adelante debido a que en el momento de la creación del modelo no se contaba con datos de quiebras para el año 2019 y, a partir del 2020, la información podría agregar demasiado ruido debido a la pandemia del covid-19. Para futuros trabajos se recomienda excluir los datos de 2020 y 2021 debido a la alta probabilidad de tener una gran cantidad de datos atípicos tras quiebras como consecuencia de las cuarentenas.

Según el Decreto 2784 de 2012, las empresas del grupo 1 son aquellas que cotizan en la bolsa de valores, las empresas de interés público, o que cuentan con una planta de personal mayor a 200 trabajadores o con activos totales superiores a 30.000 salarios mínimos mensuales legales vigentes (SMMLV) y que, adicionalmente, cumplan con cualquiera de los siguientes parámetros:

- Ser subordinada o sucursal de una compañía extranjera que aplique NIIF plenas.
- Ser subordinada o matriz de una compañía nacional que debe aplicar NIIF plenas.
- Ser matriz, asociada o negocio conjunto de una o más entidades extranjeras que aplican NIIF plenas.
- Realizar importaciones o exportaciones que representen más del 50% de las compras o de las ventas respectivamente.

Las empresas del grupo 2 poseen ingresos brutos anuales iguales o superiores a los 6000 SMMLV, tienen activos totales entre 500 y 30.000 SMMLV o cuentan con una nómina de personal entre 11 y 200 trabajadores. Se incluyen las microempresas, de igual número de ingresos, que tienen activos totales de 500 SMMLV, excluyendo la vivienda; o que cuentan con máximo 10 trabajadores.

Para la marca de procesos de insolvencia se tomó la información reportada por la Superintendencia de Sociedades de la delegatura de insolvencia donde se informa el NIT de la empresa, el nombre de la empresa, la causal de insolvencia y la fecha en la que entró a quiebra. Las que tienen activos totales de entre 500 y 30 000 SMMLV o cuentan con una nómina de personal de entre 11 y 200 trabajadores.

Se quiso implementar un filtro para excluir aquellas empresas con una edad inferior a 5 años. Sin embargo, la Superintendencia de Sociedades no publica

la fecha de constitución en la información básica disponible en el Sistema Integrado de Información Societaria (SIIS), lo que imposibilita su cálculo.

Pese a este impedimento, en el presente trabajo se asume que es probable que para cumplir los requisitos de ingresos, activos y personal solicitados para ser empresas grupo 1 o grupo 2, las empresas que publicaron su información podrían tener entre 3 o 5 años de constitución. Con base en lo anterior, se asume que trabajaremos con empresas que, de entrar a quiebra, ya tenían años de funcionamiento estable[1].

El periodo de 1995 a 2015 parece amplio, pero fue seleccionado con el fin de agregar la siguiente información a los datos:

- Una serie histórica que incluye la crisis del 98, periodos de auge de 2001 a 2007, el coletazo de la crisis de 2008 de Estados Unidos y periodos de estabilidad y crecimiento moderado.
- En el set de entrenamiento se excluyen periodos posteriores a 2016 debido a que se desea continuar con un estándar de contabilidad y poner a prueba los indicadores financieros ante un ajuste en los estándares contables. Lo anterior permite validar los resultados del modelo independiente del estándar contable, ya que lo relevante para la evaluación serán indicadores financieros.
- Para futuras extensiones de este modelo, se pueden identificar oportunidades de mejora con el uso de datos más recientes, a medida que la calidad de los datos aumente. Sin embargo, esta serie de tiempo debe ser revisada con cuidado. Los resultados a partir de 2020 deberían excluirse debido a que la coyuntura atípica de la pandemia relacionada con el covid-19 podría alterar el entrenamiento e, incluso, los resultados en sets de evaluación.
- Debido a que necesitamos conocer el estado dos años adelante, el límite es 2017 porque esto nos llevará a que en 2019 ya se sabrá si presentaron quiebra o no. Para el tiempo de esta tesis no están disponibles los procesos para 2020 y, de estar disponibles, deberían ser excluidos por la atipicidad de la pandemia.

1 Los estados financieros están disponibles para la descarga en el SIIS en el siguiente enlace: https://siis.ia.supersociedades.gov.co/#/, Los Procesos de Insolvencia están disponibles para la descarga en el siguiente enlace: https://www.supersociedades.gov.co/delegatura_insolvencia/Paginas/publicaciones.aspx

### 3.1 Preparación de los datos

La preparación de los datos y las diferentes versiones del algoritmo fueron desarrolladas en Alteryx, un *software* de analítica que reemplaza los códigos de programación por una interfaz gráfica basada en "Drag and Drop". Este *software* permite ver de manera visual el flujo de información en donde cada componente que se lleva al escritorio es un paso en la trasformación y el procesamiento de los datos. Si bien se usó Alteryx para facilitar el trabajo, es completamente replicable en R.

En este trabajo se crearon dos sets de datos principales de los cuales se derivan los tres sets para entrenar cada modelo. El primer set es el de entrenamiento y validación con datos desde 1995 a 2015. El segundo set tiene los datos empresariales de 2016 para la evaluación del modelo.

### 3.2 Estructura del set de datos de estados financieros– entrenamiento y evaluación

Al unir todos los sets de datos de empresas del grupo 1 y del grupo 2 para cada uno de los años, desde 1995 a 2015 y el 2016, para la evaluación se obtuvo un total de 266.357 registros para el set de entrenamiento y 11.680 para el set de evaluación. Cada columna del set de datos cuenta con el NIT de la empresa, el año al que corresponden los estados financieros y las cuentas de los estados financieros listadas a continuación.

ODEON, ISSN: 1794-1113, E-ISSN: 2346-2140, N° 27, julio-diciembre de 2024, pp. 209-254

Tabla 3. Campos del set de datos de entrenamiento y evaluación

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| NIT | 1320 CUENTAS POR COBRAR A VINCU. ECONÓMICOS | 1450 TERRENOS | 1370 PRÉSTAMOS A PARTICULARES | 1995 DE OTROS ACTIVOS | 25 OBLIGACIONES LABORALES CORTO PLAZO | 2305 CUENTAS CORRIENTES COMERCIALES (LP) | 2810 DEPOSITOS RECIBIDOS (LP) | 3105 CAPITAL SUSCRITO Y PAGADO |
| YEAR | 1323 CUENTAS POR COBRAR A DIRECTORES | 1455 MATERIALES REPUESTOS Y ACCESORIOS | 1380 DEUDORES VARIOS (LP) | 19 SUBTOTAL VALORIZACIONES | 2605 PARA COSTOS Y GASTOS | 2310 A CASA MATRIZ (LP) | 2815 INGRESOS RECIBIDOS PARA TERCEROS (LP) | 3115 APORTES SOCIALES |
| INGRESOS OPERACIONALES | 1325 CUENTAS POR COBRAR A SOCIOS Y ACCIONISTAS | 1460 ENVASES Y EMPAQUES | 1385 DERECHOS DE RECOMPRA CARTERA NEGOCIADA (LP) | 0 TOTAL ACTIVO NO CORRIENTE | 2610 PARA OBLIGACIONES LABORALES | 2315 A COMPAÑÍAS VINCULADAS (LP) | 2820 CUENTAS DE OPERACIÓN CONJUNTA | 3120 CAPITAL ASIGNADO |
| COSTO DE VENTAS | 1328 APORTES POR COBRAR | 1465 INVENTARIOS EN TRÁNSITO | 1390 DEUDAS DE DIFICIL COBRO (LP) | 0 TOTAL ACTIVO | 2615 PARA OBLIGACIONES FISCALES | 2320 A CONTRATISTAS (LP) | 2825 RETENCIÓN A TERCEROS SOBRE CONTRATOS (LP) | 3125 INV SUPLEMEN AL CAPITAL ASIGNADO |
| UTILIDAD BRUTA | 1330 ANTICIPOS Y AVANCES | 1499 PROVISIONES | 1399 PROVISIONES (LP) | 81 DERECHOS CONTINGENTES | 2620 PENSIONES DE JUBILACIÓN | 2335 COSTOS Y GASTOS X PAGAR | 2835 ACREEDORES DEL SISTEMA (LP) | 3130 CAPITAL DE PERSONAS NATURALES |
| GASTOS ADMINISTRACIÓN | 1332 CTAS DE OPERACIÓN CONJUNTA | 14 SUBTOTAL INVENTARIOS | 13 SUBTOTAL DEUDORES LARGO PLAZO | 82 DEUDORAS FISCALES | 2625 PARA OBRAS DE URBANISMO | 2345 ACREEDORES OFICIALES (LP) | 2840 CUENTAS EN PARTICIPACIÓN (LP) | 3135 APORTES DEL ESTADO |
| GASTOS VENTAS | 1335 DEPÓSITOS | 1705 GASTOS PAGADOS X ANTICIPADO | 15 PROPIEDADES PLANTA Y EQUIPO NETO | 83 DEUDORAS DE CONTROL | 2630 PARA MANTENIMIENTO Y REPARACIONES | 2350 REGALÍAS X PAGAR | 2895 DIVERSOS (LP) | 3140 FONDO SOCIAL |

| UTILIDAD OPERACIONAL | 1340 PROMESAS DE COMPRAVENTA | 1710 CARGOS DIFERIDOS | 1605 CRÉDITO MERCANTIL | 9 CUENTAS DE ORDEN ACREEDORES POR CONTRA | 2635 PARA CONTINGENCIAS | 2355 DEUDAS CON ACCIONISTAS O SOCIOS (LP) | 28 SUBTOTAL OTROS PASIVOS LARGO PLAZO | 31 SUBTOTAL CAPITAL SOCIAL |
|---|---|---|---|---|---|---|---|---|
| INGRESOS NO OPERACIONALES | 1345 INGRESOS POR COBRAR | 1715 COSTOS DE EXPLORACIÓN X AMORTIZAR | 1610 MARCAS | 21 OBLIGACIONES FINANCIERAS (CP) | 2640 PARA OBLIGACIONES DE GARANTÍAS | 2357 DEUDAS CON DIRECTORES (LP) | 2905 BONOS EN CIRCULACIÓN (LP) | 3205 PRIMA EN COLOC. ACC CUOTAS O PARTES DE INT. S |
| GASTOS NO OPERACIONALES | 1350 RETENCIÓN SOBRE CONTRATOS (CP) | 1720 COSTOS DE EXPLOTACIÓN Y DESARROLLO | 1615 PATENTES | 22 PROVEEDORES | 2695 PROVISIONES DIVERSAS | 2360 DIVIDENDOS O PARTIC. X PAGAR (LP) | 2910 BONOS OBLIGATORIAMENTE CONVERTIB. ACCION (LP) | 3210 DONACIONES |
| UTILIDAD NETA ANTES DE IMPUESTOS | 1355 ANTIC IMPTOS Y CONTRIB O SALDOS A FAVOR | 1730 CARGOS POR CORREC. MONET. DIFERIDA | 1620 CONCESIONES Y FRANQUICIAS | 2305 CUENTAS CORRIENTES COMERCIALES (CP) | 26 SUBTOTAL PASIVOS ESTIMADOS Y PROVISIONES | 2375 CUOTAS POR DEVOLVER (LP) | 2915 PAPELES COMERCIALES (LP) | 3215 CRÉDITO MERCANTIL |
| IMPUESTO DE RENTA | 1360 RECLAMACIONES (CP) | 1798 AMORTIZACIÓN ACUMULADA | 1625 DERECHOS | 2310 A CASA MATRIZ (CP) | 27 DIFERIDOS CORTO PLAZO | 2380 ACREEDORES VARIOS (LP) | 2920 BONOS PENSIONALES (LP) | 3220 KNOW HOW |
| UTILIDAD NETA | 1365 CUENTAS X COBRAR A TRABAJADORES (CP) | 17 SUBTOTAL DIFERIDO | 1630 KNOW HOW | 2315 A COMPAÑÍAS VINCULADAS (CP) | 2805 ANTICIPOS Y AVANCES RECIBIDOS (CP) | 23 SUBTOTAL CUENTAS POR PAGAR LARGO PLAZO | 2925 TÍTULOS PENSIONALES (LP) | 3225 SUPERÁVIT MÉTODO E PARTICIPACIÓN |
| 47 AJUSTES POR INFLACIÓN | 1370 PRÉSTAMOS A PARTICULARES | 0 TOTAL ACTIVO CORRIENTE | 1635 LICENCIAS | 2320 A CONTRATISTAS (CP) | 2810 DEPÓSITOS RECIBIDOS (CP) | 25 OBLIGACIONES LABORALES LARGO PLAZO | 29 SUBTOTAL BONOS Y PAPELES COMERCIALES (LP) | 32 SUBTOTAL SUPERÁVIT DE CAPITAL |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 530520 _ INTERESES | 1380 DEUDORES VARIOS (CP) | 12 INVERSIONES LP | 1698 AMORTIZACIÓN ACUMULADA | 2330 ÓRDENES DE COMPRA X UTILIZAR | 2815 INGRESOS RECIBIDOS PARA TERCEROS (CP) | 2605 PARA COSTOS Y GASTOS (LP) | 0 TOTAL PASIVO NO CORRIENTE | 33 RESERVAS |
| RAZÓN SOCIAL | 1385 DERECHOS RECOMPRA CARTERA NEGOCIADA (CP) | 1305 CLIENTES (LP) | 1699 PROVISIONES | 2335 COSTOS Y GASTOS X PAGAR | 2820 CUENTAS DE OPERACIÓN CONJUNTA | 2610 PARA OBLIGACIONES LABORALES (LP) | 0 TOTAL PASIVO | 34 REVALORIZACIÓN DEL PATRIMONIO |
| CIUDAD | 1390 DEUDAS DE DIFÍCIL COBRO (CP) | 1310 CUENTAS CORRIENTES COMERCIALES (LP) | 16 SUBTOTAL INTANGIBLES | 2340 INSTALAMENTOS X PAGAR | 2825 RETENCIÓN A TERCEROS SOBRE CONTRATOS (CP) | 2615 PARA OBLIGACIONES FISCALES (LP) | | 35 DIVIDEN O PARTC. DECRETADAS EN ACC.O CUOTAS |
| DPTO | 1399 PROVISIONES (CP) | 1315 CUENTAS X COBRAR A CASA MATRIZ (LP) | 1705 GASTOS PAGADOS X ANTICIPADO | 2345 ACREEDORES OFICIALES (CP) | 2830 EMBARGOS JUDICIALES | 2620 PENSIONES DE JUBILACIÓN (LP) | | 36 RESULTADOS DEL EJERCICIO |
| CIIU | 13 SUBTOTAL DEUDORES CORTO PLAZO | 1320 CUENTAS POR COBRAR A VINCULADOS ECONÓMICOS | 1710 CARGOS DIFERIDOS | 2350 REGALÍAS X PAGAR | 2835 ACREEDORES DEL SISTEMA (ANEXO 18) | 2625 PARA OBRAS DE URBANISMO (LP) | | 37 RESULTADOS DE EJERCICIOS ANTERIORES |
| DESCRIPCIÓN SECTOR | 1405 MATERIAS PRIMAS | 1323 CUENTAS POR COBRAR A DIRECTORES | 1715 COSTOS DE EXPLORACIÓN X AMORTIZAR | 2355 DEUDAS CON ACCIONISTAS O SOCIOS (CP) | 2840 CUENTAS EN PARTICIPACIÓN | 2635 PARA CONTINGENCIAS (LP) | | 38 SUPERÁVIT POR VALORIZACIONES |
| 1105 CAJA | 1410 PRODUCTOS EN PROCESO | 1325 CUENTAS X COBRAR A SOCIOS Y ACCIONISTAS (LP) | 1720 COSTOS DE EXPLOTACIÓN Y DESARROLLO | 2357 DEUDAS CON DIRECTORES | 2895 DIVERSOS (CP) | 2640 PARA OBLIGACIONES DE GARANTÍAS (LP) | | 0 TOTAL PATRIMONIO |

ODEON, ISSN: 1794-1113, E-ISSN: 2346-2140, N° 27, julio-diciembre de 2024, pp. 209-254

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1110 BANCOS | 1415 OBRAS DE CONSTRUC EN CURSO | 1330 ANTICIPOS Y AVANCES (LP) | 1730 CARGOS POR CORREC. MONET. DIFERIDA | 2360 DIVIDENDOS O PARTIC. X PAGAR | 28 SUBTOTAL OTROS PASIVOS CORTO PLAZO | 2695 PROVISIONES DIVERSAS (LP) | | 0 TOTAL PASIVO Y PATRIMONIO |
| 1115 REMESAS EN TRÁNSITO | 1417 OBRAS DE URBANISMO | 1332 CUENTAS DE OPERACIÓN CONJUNTA | 1798 AMORTIZACIÓN ACUMULADA | 2365 RETENCIÓN EN LA FUENTE | 2905 BONOS EN CIRCULACIÓN | 26 SUBTOTAL PASIVOS ESTIMADOS Y PROVISIONES (LP) | | 91 RESPONSABILIDADES CONTINGENTES |
| 1120 CUENTAS DE AHORRO | 1420 CONTRATOS EN EJECUCIÓN | 1335 DEPÓSITOS (LP) | 17 SUBTOTAL DIFERIDOS | 2367 IMPUESTO A LAS VENTAS RETENIDO | 2910 BONOS OBLIGATORIAMENTE CONVERTIBLES EN ACCIÓN | 2705 INGRESOS RECIBIDOS X ANTICIPADO (LP) | | 92 ACREEDORAS FISCALES |
| 1125 FONDOS | 1425 CULTIVOS EN DESARROLLO | 1340 PROMESA DE COMPRAVENTA (LP) | 1805 BIENES DE ARTE Y CULTURA | 2368 IMPUESTO DE INDUSTRIA Y COMERCIO RETENIDO | 2915 PAPELES COMERCIALES | 2710 ABONOS DIFERIDOS (LP) | | 93 ACREEDORAS DE CONTROL |
| 11 SUBTOTAL DISPONIBLE | 1428 PLANTACIONES AGRÍCOLAS | 1345 INGRESOS POR COBRAR | 1895 DIVERSOS | 2370 RETENCIONES Y APORTES DE NÓMINA | 2920 BONOS PENSIONALES | 2715 UTILIDAD DIFERIDA EN VTAS A PLAZOS (LP) | | 8 CUENTAS DE ORDEN DEUDORAS POR CONTRA |
| 12 INVERSIONES | 1430 PRODUCTOS TERMINADOS | 1350 RETENCIÓN SOBRE CONTRATOS (LP) | 1899 PROVISIONES | 2375 CUOTAS POR DEVOLVER CP | 2925 TÍTULOS PENSIONALES | 2720 CRÉDITO X CORREC. MONETARIA DIFERIDA (LP) | | 22 PROVEEDORES (ANEXO 10) (LP) |
| 1305 CLIENTES | 1435 MCIAS NO FABRICADAS X LA EMPRESA | 1355 ANTICIPO DE IMPTOS. Y CONTRIB. O SALDOS A FAV | 18 SUBTOTAL OTROS ACTIVOS | 2380 ACREEDORES VARIOS (CP) | 29 SUBTOTAL BONOS Y PAPELES COMERCIALES | 2725 IMPUESTOS DIFERIDOS (LP) | | 24 IMPUESTOS GRAVÁMENES Y TASAS (LP) |



ODEON, ISSN: 1794-1113, E-ISSN: 2346-2140, N° 27, julio-diciembre de 2024, pp. 209-254

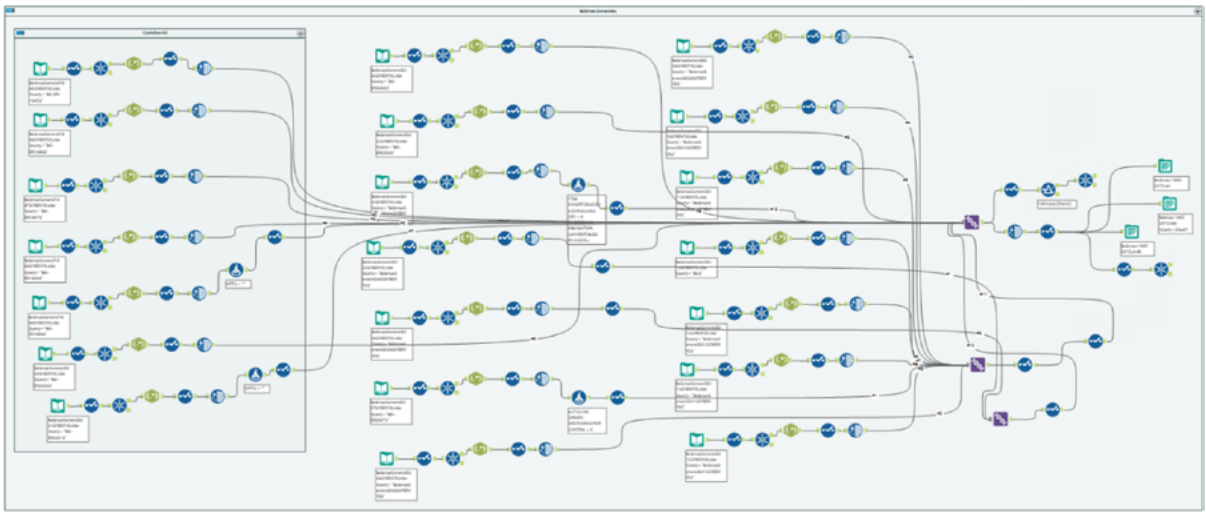
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1310 CUENTAS CORRIENTES COMERCIALES | 1440 BIENES RAÍCES PARA LA VENTA | 1360 RECLAMACIONES (LP) | 1905 DE INVERSIONES | 23 SUBTOTAL CUENTAS POR PAGAR CORTO PLAZO | 0 TOTAL PASIVO CORRIENTE | 27 SUBTOTAL DIFERIDOS LARGO PLAZO | | 3705 UTILIDADES ACUMULADAS |
| 1315 CUENTAS POR COBRAR A CASA MATRIZ | 1445 SEMOVIENTES | 1365 CUENTAS POR COBRAR A TRABAJADORES (LP) | 1910 DE PROPIEDADES PLANTA Y EQUIPO | 24 IMPUESTOS GRAVÁMENES Y TASAS | 21 OBLIGACIONES FINANCIERAS (LP) | 2805 AVANCES Y ANTICIPOS RECIBIDOS (LP) | | 3710 PÉRDIDAS ACUMULADAS |

Fuente: Reportes estados financieros Superintendencia de Sociedades (1995-2016).

ODEON, ISSN: 1794-1113, E-ISSN: 2346-2140, N° 27, julio-diciembre de 2024, pp. 209-254

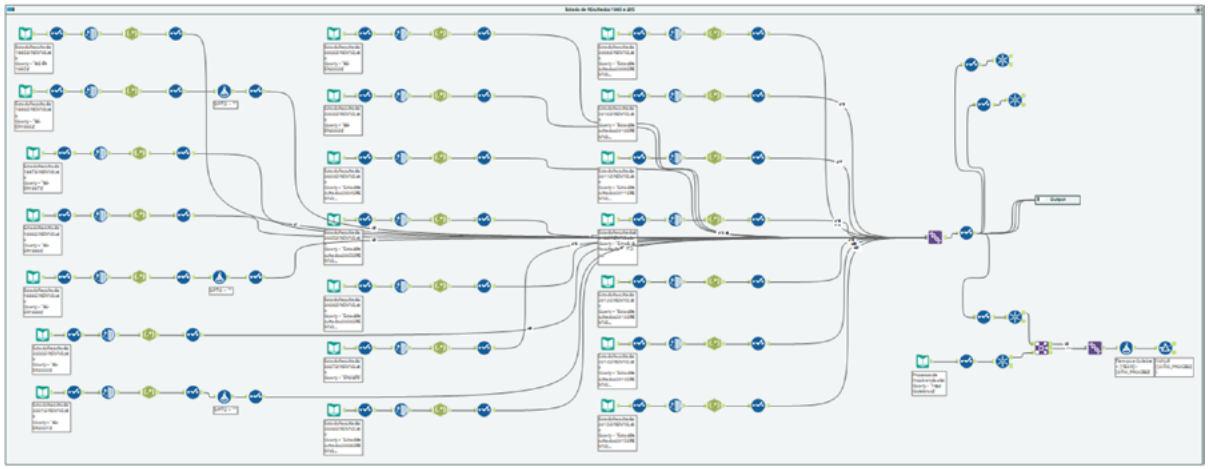
Esta base es la fuente para calcular los indicadores que se usarán para entrenar y evaluar el set de datos para el modelo Z-Score, El modelo Altman, Barboza, Kimura y el propuesto en el presente trabajo.

Figura 1. Unión de los archivos de Balance general de la SuperSociedades, 1995-2015 (captura de pantalla Alteryx)



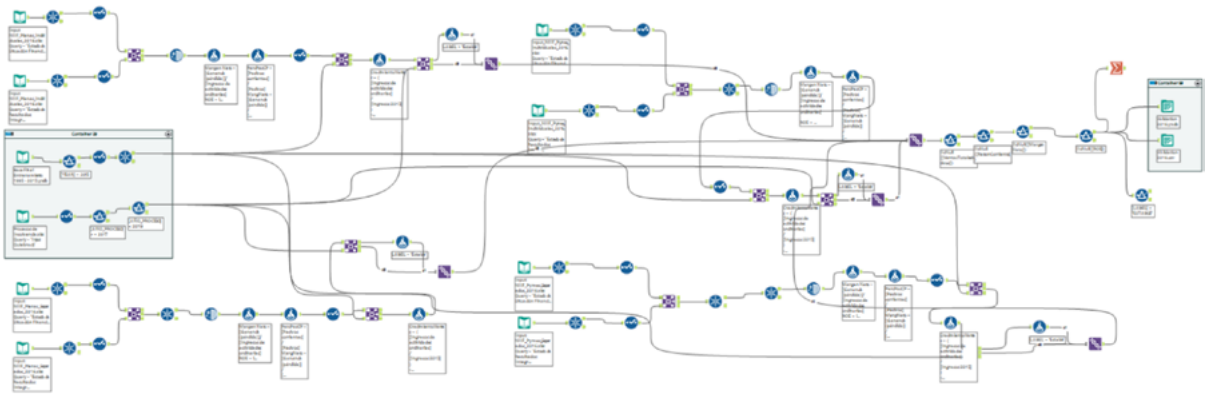
Fuente: elaboración propia.

Figura 2. Unión de los archivos de Estado de resultados de la SuperSociedades, 1995-2015 (captura pantalla Alteryx)



Fuente: elaboración propia.

Figura 3. Unión de archivos para crear el set de evaluación, 2016 (captura de pantalla Alteryx)



Fuente: elaboración propia.

## 3.1 Estructura del set de datos de quiebra–entrenamiento y evaluación

La fuente de datos para obtener la marca de quiebra o estable y completar el set de entrenamiento y evaluación contiene el NIT de la empresa, el proceso de liquidación iniciado ante la Superintendencia Financiera y el año en que inició el proceso.

Tabla 4. Ejemplo de la estructura del ser de datos de la marca Estable/Quiebra

| NIT | Proceso | Año Proceso |
|---|---|---|
| 81110 | Liquidación por adjudicación | 2016 |
| 169044 | Reorganización | 2018 |
| 212759 | Reorganización en ejecución | 2012 |
| 223513 | Reorganización | 2017 |
| 322549 | Reorganización en ejecución | 2017 |
| 405741 | Reorganización en ejecución | 2014 |
| 428616 | Reorganización | 2019 |
| 428618 | Reorganización | 2019 |

Fuente: elaboración propia.

Esta base de datos será fusionada con la de indicadores financieros a través del campo NIT, agregándole la fecha en que se inició el proceso de Quiebra. Una vez se obtiene esta marca se puede comparar el año de los estados financieros con el inicio de proceso de quiebra para obtener la marca de Quiebra o Estables.

### 3.2 Creación de la marca Quiebra/Estable

Con el año de los estados financieros y el año de quiebra en una misma línea, se identificará si la empresa permanecerá estable durante los próximos dos años o entrará en quiebra. Para lograrlo, al año de inicio de quiebra se le restará el año de los estados financieros y aquellos que den como resultado 2 o menos años de diferencia se les asignará la marca de Quiebra, a las demás compañías se les asignará la marca de Estable.

En el set de datos de evaluación (los de 2016), la marca Quiebra se les asignará a aquellas compañías que se hayan acogido a reorganización o hayan sido liquidadas por orden de la SuperSociedades para los años 2017 y 2018.

Con la marca de Quiebra y Estable se excluirán los estados financieros de empresas que entraron en quiebra hace 2 años pero que continúan reportando sus cifras a la Superintendencia de Sociedades. Lo anterior quiere decir que si tenemos una empresa que entró en quiebra en 2010, se excluirán los periodos del 2012 en adelante. Esto nos dejará estados financieros de empresas que reportaron quiebra al año o a los dos años.

### 3.3 Estructura del set de datos de entrenamiento y evaluación

Con base en el set de datos de estados financieros se calcularon los indicadores necesarios para el modelo y se excluyeron todas las columnas que no se necesitan para el entrenamiento y evaluación. Los sets finales, entrenamiento y evaluación, incluyen las siguientes columnas de información.

Tabla 5. Estructura del set de datos de entrenamiento y evaluación

| NIT | Número de Identificación de la Empresa |
|---|---|
| YEAR | Año al que corresponden los estados financieros e Indicadores |
| CapTrabActTot | Capital de Trabajo / Activos Totales |
| UtRetActTot | Utilidades retenidas / Activos Totales |
| UTNetaToAct | Utilidad Neta/ total de Activos |
| VentasTotalesActivos | Ventas Totales / Total Activos |

| NIT | Número de Identificación de la Empresa |
|---|---|
| ActPAs | Total Activo / Total Pasivo |
| ROE | Retorno sobre el patrimonio |
| CrecimientoVentas | Crecimiento en ventas |
| CrecimientoActivo | Crecimiento de los activos |
| PercPasCP | Pasivo CP / Total Pasivo |
| MargNeto | Margen Neto |
| Razón Corriente | Razón Corriente |
| LABEL | Marca de Quiebra o Estable |

Fuente: elaboración propia.

Con la fusión de bases de datos se tiene toda la información disponible para entrenar y evaluar los modelos, no obstante, se debe evaluar la calidad de los datos y hacer una exploración de los mismos para tener una base de datos limpia que arroje un modelo útil para la aplicación en la vida real.

## 3.6 Limpieza de los datos

Con el fin de excluir de la base de datos los registros con errores y datos atípicos, se realizó un proceso de limpieza de datos con algoritmos que permitieran preparar cada set sin necesidad de cambios manuales. Este proceso permite replicar esta tesis con información de años futuros sin necesidad de cambios en el código realizado en R.

### 3.6.1. Limpieza del set de entrenamiento 1995-2015–balance general y estado de resultados

- A cada archivo de estados financieros se le asignó un nombre con el número del año de cuatro (4) dígitos.
- Se usó RegEx con la expresión "\d{4}" para extraer del nombre del archivo el año al cual pertenece cada registro. El año de cada archivo se almacenó en la columna YEAR. Este año servirá para identificar la fecha de los estados financieros y para calcular la diferencia con el año de quiebra, si esta existió.
- Para los datos de Balance General de 1999 y 2001 se tuvo que agregar la columna que contiene la información de departamento de constitución de

la compañía "DPTO" debido a que no estaba disponible. Este ajuste permite tener las mismas columnas en todos los años y poder unir en un solo set de datos todos los años de estados financieros. Este ajuste para mantener las buenas prácticas en el momento de unir dos o más sets de datos.

- Para el balance general del año 2004 se agregaron las columnas 1798 AMORTIZACIÓN ACUMULADA (CP), 2910 BONOS OBLIGATOR. CONVERTIBLES EN ACCIONES (CP) y 3135 APORTES DEL ESTADO.
- Para el balance general de 2007 se agregó la columna 8 CTAS DE ORDEN DEUDORAS POR EL CONTRARIO.
- Tanto para balance general como para el estado de resultados se reemplazaron los valores nulos por CEROS.
- Para Estados de Resultados de 1996, 1999 y 2001 se agregó en blanco la columna departamento y se eliminaron duplicados.

### 3.6.2. Limpieza set de procesos de insolvencia

- Se excluyeron duplicados.
- Se encontraron algunos *label* con espacios antes o después de la palabra Reorganización por lo que se usó la función TRIM para eliminar espacios al principio y al final de las columnas. Aunque esta modificación no afecta la marca de Quiebra y Estable es mejor mantener los datos lo más limpios posible.

### 3.6.3. Unión de estados financieros por año y por grupo y creación de la marca Estable/Quiebra

Con los tres sets de datos limpios se procedió a unirlos en un solo gran set de datos:

- Agregar a los datos de los estados de resultados con un LEFT JOIN los datos de balance. Se unió con base en NIT y año.
- Se agregó el año de quiebra uniendo el dataset de solvencia con el anterior usando la columna NIT.
- Se calculó el tiempo a la quiebra como la resta entre el año del estado de resultados y el año del reporte de quiebra informado por la Superintendencia de Sociedades.
- Se excluyen los estados financieros de empresas que declararon quiebras y se reportaron después del suceso. Tiempo a quiebra <= 0 o mayor

- Se calcula la marca de quiebra con el siguiente condicional.

    - if [Tiempo a Quiebra] >= -2 then ‘Quiebra’ else ‘Estable’ endif

- Se filtraron los registros que no tenían ingresos usando ingresos mayores a 0.
- Resumen de marcas.

    - Estable: no reportó quiebra.
    - Quiebra: reportó quiebra al año o dos años siguientes.

## 3.7 Cálculo de los indicadores financieros

Con la base de datos resultante de la primera fase de la limpieza de datos se procedió a calcular los indicadores financieros necesarios para entrenar los modelos. A continuación, se presentan los cálculos con los nombres de campos utilizados en el procesamiento de la información.

Tabla 6. Cálculo de indicadores financieros del modelo

| NIT | Número de Identificación de la Empresa |
|---|---|
| YEAR | Año al que corresponden los estados financieros e Indicadores |
| CapTrabActTot | Capital de Trabajo / Activos totales |
| UtRetActTot | Utilidades retenidas / Activos Totales |
| UTNetaToAct | Utilidad Neta/ Total de Activos |
| VentasTotalesActivos | Ventas Totales / Total Activos |
| ActPas | Total Activo / Total Pasivo |
| ROE | Retorno sobre el patrimonio |
| CrecimientoVentas | Crecimiento en ventas |
| CrecimientoActivo | Crecimiento de los activos |
| PercPasCP | Pasivo CP / Total Pasivo |
| MargNeto | Margen Neto |
| Razón Corriente | Razón Corriente |
| LABEL | Marca de Quiebra o Estable |

Fuente: elaboración propia.

Antes de empezar con los cálculos, debido a la forma en que se trataron los datos, algunos campos podrían estar en texto en vez de número, por lo que se convirtieron los campos nulos en Cero y se forzó a que los valores tuvieran un formato numérico.

## 3.8 Preparación del set de evaluación–estados financieros en 2016

En 2016 se unieron los siguientes archivos:

- NIIF plenas individuales 2016 – Estado de situación financiera, Estado de Resultados.
- NIIF plenas Separados 2016 – Estado de situación financiera, Estado de Resultados.
- NIIF Pymes individuales 2016 – Estado de situación financiera, Estado de Resultados.
- NIIF Pymes Separados 2016 – Estado de situación financiera, Estado de Resultados.

A esta unión se le realizaron los mismos cálculos y la misma limpieza descrita para los estados financieros de 1995-2015. Adicional, a este set de datos se trajo para cada registro el valor del activo y los ingresos del año 2015 para calcular las variaciones de los activos y de los ingresos.

Por último, se realizó una limpieza del set de datos adicional, que evitaba tener registros incompletos. Se realizó el filtro de razón corriente no nula, lo que excluyó 264 registros del set de entrenamiento. Se eliminó la utilidad operativa por permanecer en ceros en muchas ocasiones; esta situación explica por qué se cambió la utilidad operativa por la utilidad neta en las variables del modelo.

### 3.8.1 Resultado después de la limpieza

- El set de entrenamiento se compone de 496 quiebras y 265.865 registros estables.
- El set de datos de evaluación con datos del 2016 se compone de 11.487 registros estables y 202 quiebras.

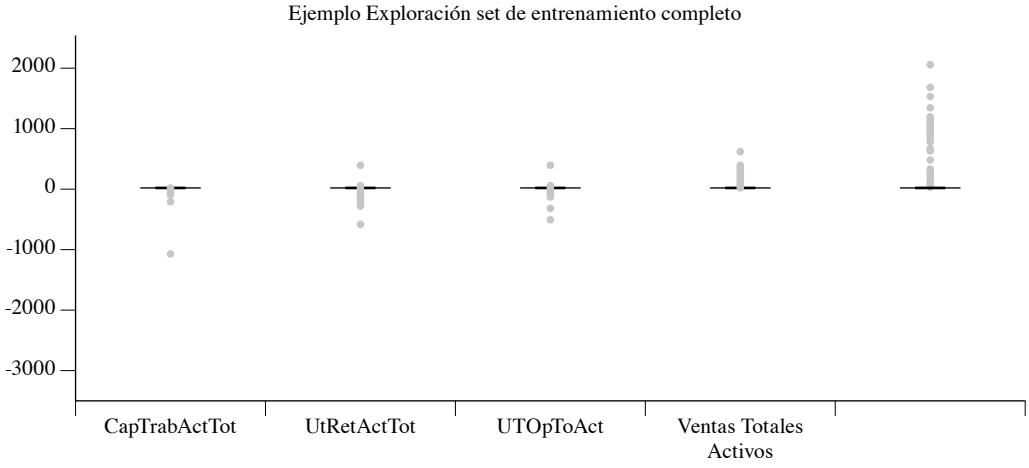
## 3.9 Rebalanceo de los sets de datos

Al revisar los resultados de la limpieza, se puede observar que la cantidad de registros estables frente a quiebras es ampliamente superior. Si se entrena un modelo de árboles de decisión o cualquier modelo de *machine learning* con este desbalance, la probabilidad de clasificar una empresa como estable aumenta frente a lo que sería una empresa en quiebra.

Para evitar este problema se debe balancear el set de datos. El objetivo con el balance es dejar un set de datos lo más cercano a 50% Quiebra, 50% Estable, siendo el mínimo aceptable 25% Quiebra y 75% Estable. Con el fin de lograrlos se seleccionará aleatoriamente una muestra de estables y se duplicarán registros de quiebra para aumentar su participación y que el modelo entienda mejor esta marca. Se decidió 40%/60% luego de varias pruebas en la precisión de los modelos. Para replicar el experimento en cada selección se estableció un valor de semilla fijo.

## 3.10 Revisión gráfica de los datos

Figura 4. SEQ Figura * ARABIC \s 1 4 Ejemplo proceso de exploración de Entrenamiento




Fuente: elaboración propia.

Luego del rebalanceo de la base de datos, se requiere la inspección visual para determinar las distribuciones de probabilidad de los datos y la presencia de datos atípicos o *outliers*.

La inspección por *box-plots* mostró la presencia de una gran cantidad de *outliers* y datos extraños, posiblemente como resultado de un mal *input* por parte de las empresas.

Para lidiar con estos datos atípicos se realizaron los siguientes filtros teniendo en la cuenta niveles del *box-plot*: se filtraron las bases de entrenamiento y de evaluación por los valores que estaban por fuera de los rangos intercuartiles de los bigotes de cada variable.

Los filtros aplicados para remover los datos atípicos se tomaron con base en los rangos intercuartiles de los diagramas de cajas o *box-plots,* se buscó retirar los datos estadísticamente atípicos, pese a que algunos rangos parecen demasiado amplios para los estándares del análisis financiero. Si bien permanecen algunas empresas con variaciones o indicadores bastante buenos o malos, estas se mantuvieron con el fin de tener la mayor cantidad de casos posibles dentro de lo que estadísticamente se considera en un rango aceptable.

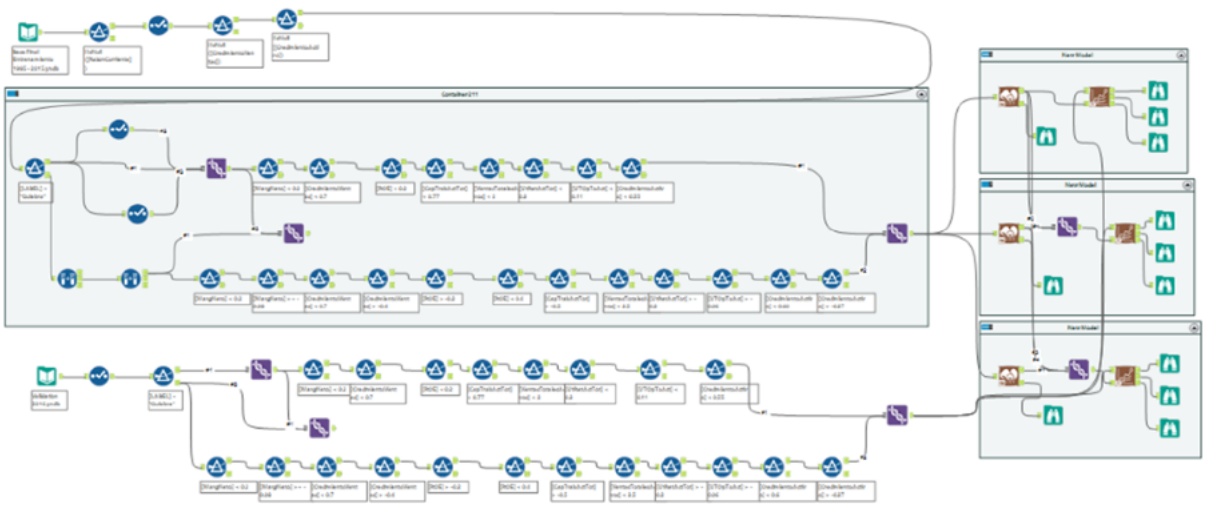
- **Filtros para datos con marca Estable**

  - Margen neto menor al 20%. Estos límites no afectarán los resultados, ya que márgenes superiores al 20% no reportaron empresas con quiebra probable y con margen neto inferior al 8% presenta resultados mixtos.
  - Crecimiento de ventas menor al 70%. Aunque son pocas las empresas con estos niveles de crecimiento, este punto de corte se tomó con base en los rangos intercuartiles de los *box-plot*. El cambio en este límite no afectó los resultados debido a la alta probabilidad de no quiebra de empresas con un crecimiento sostenido en ventas.
  - ROE menor al 20%, aquellos por encima de este nivel son todos estables.
  - Capital de trabajo / Activos Totales menor al 77%.
  - Ventas Totales / Activos Totales menor a 2.
  - Utilidades Retenidas / Activos Totales Menor al 30%.
  - Ut Operativa / Activos – en este caso es la utilidad neta menor al 11%.
  - Crecimiento de los activos menor a 55%.

- **Filtros para datos con marca Quiebra**

  - Margen neto menor al 20%, Margen mayor o igual a 8%. Estos límites no afectarán los resultados, ya que márgenes superiores al 20% no se reportaron empresas con quiebra probable y con margen neto inferior al 8% presenta resultados mixtos.
  - Crecimiento Ventas menor al 70%.
  - Empresas con caída en ventas superiores a un 40% o en términos del condicional aplicado, Crecimiento Ventas mayor a -40%.
  - ROE menor al 40%.
  - ROE mayor a -20%.
  - Capital de trabajo / Activos Totales mayor a -50%.
  - Ventas totales / Activos Totales menor a 3,5.
  - Utilidades Retenidas / Activos Totales mayor a -20%
  - Ut Operativa / Activos–en este caso, es la utilidad neta mayor a -6%
  - Crecimiento de los activos menor a 60%
  - Disminución de activos que no sobrepase el 37% o en términos del condicional aplicado, crecimiento de los activos mayor a -37%.

Figura 5. Filtro de datos atípicos y entrenamiento de los modelos (captura de pantalla Alteryx)



Fuente: elaboración propia.

Después de los ajustes para balancear los datos y excluir los registros atípicos, se obtuvo para el set de datos de entrenamiento y validación un total de 1473 registros de quiebra y se seleccionaron aleatoriamente 2125 registros de empresas estables. Con esta muestra se seleccionaron los sets de entrenamiento y validación con una distribución aleatoria debido a que la cantidad de registros de quiebra fue incrementada por medio de duplicar registros aleatorios con el fin de mejorar el balance de los datos asumiendo el riesgo de sobreajuste como una oportunidad de identificar alertas de quiebra temprana.

Para asegurar la efectividad del modelo, se sometió el set de datos de evaluación (datos de 2016) al mismo proceso de limpieza y filtro de datos atípicos, pero no al rebalanceo. El objetivo de omitir el rebalanceo era probar el algoritmo en un escenario lo más cercano posible a la realidad donde la quiebra tiende a ser menos frecuente que la estabilidad y es un estado en el que se espera una mayor probabilidad de error en las predicciones.

Tabla 7. Distribución en set de datos de entrenamiento, validación y evaluación

| | Quiebra | Estable | Total |
|---|---|---|---|
| Set entrenamiento y validación | 1473 | 2125 | 3579 |
| Set entrenamiento | 586 | 584 | 1170 |
| Set validación | 1171 | 1167 | 1080 |
| Set evaluación | 194 | 5885 | 6016 |

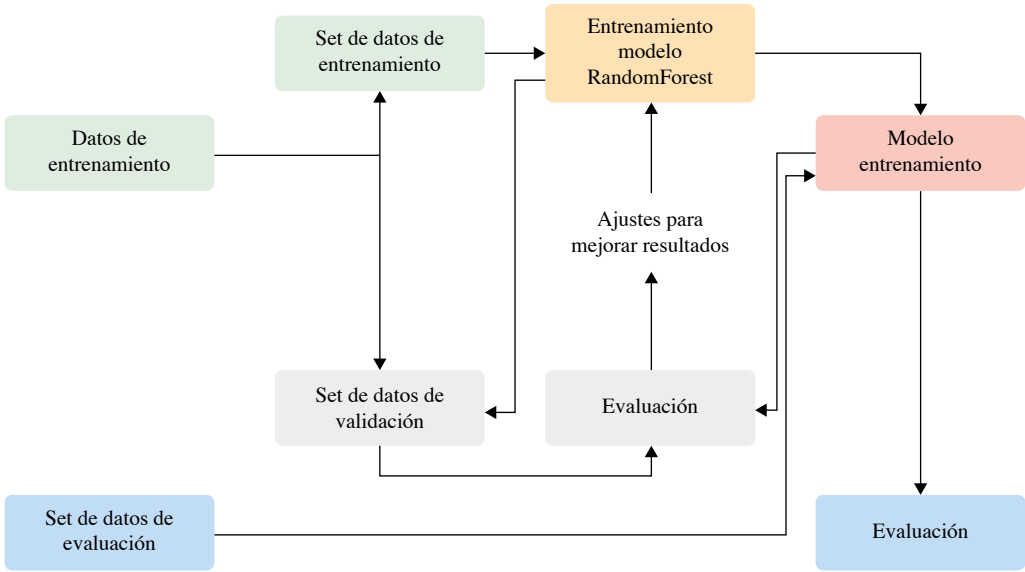
La sumatoria de registros del set de entrenamiento por separado
es mayor al total de registros consecuencia del proceso de rebalanceo de los sets de datos.

Fuente: elaboración propia.

## 3.11 Entrenamiento y validación de los modelos

Con el set de datos de entrenamiento limpio y sin datos atípicos, se entrenaron los tres modelos de Random Forest, en cada uno seleccionadas las variables que los componen. Para el entrenamiento se usó el algoritmo con 100 árboles, mínimo 5 datos por cada nodo del árbol y para cada árbol creado se seleccionó aleatoriamente el 50% de los datos. Para cada propuesta, la del Z-Score, la de Altman, Barboza, Kimura y la del presente trabajo se aplicará el mismo proceso en el mismo set de datos, pero filtrando las variables a usar para cada caso.

Figura 6. Proceso de entrenamiento y evaluación para obtener el modelo Random Forest




Fuente: Trappenberg (2020).

### 3.11.1 El modelo de Alteryx y el paquete Random Forest en R

La mayoría de los módulos de analítica de Alteryx se basan en paquetes o librerías de R, software libre desarrollado para estadística, por esto, lo presentado en este trabajo se puede replicar completamente en R con los mismos resultados siempre y cuando se usen los mismos valores de semilla en aquellos procesos que involucran algún componente aleatorio.

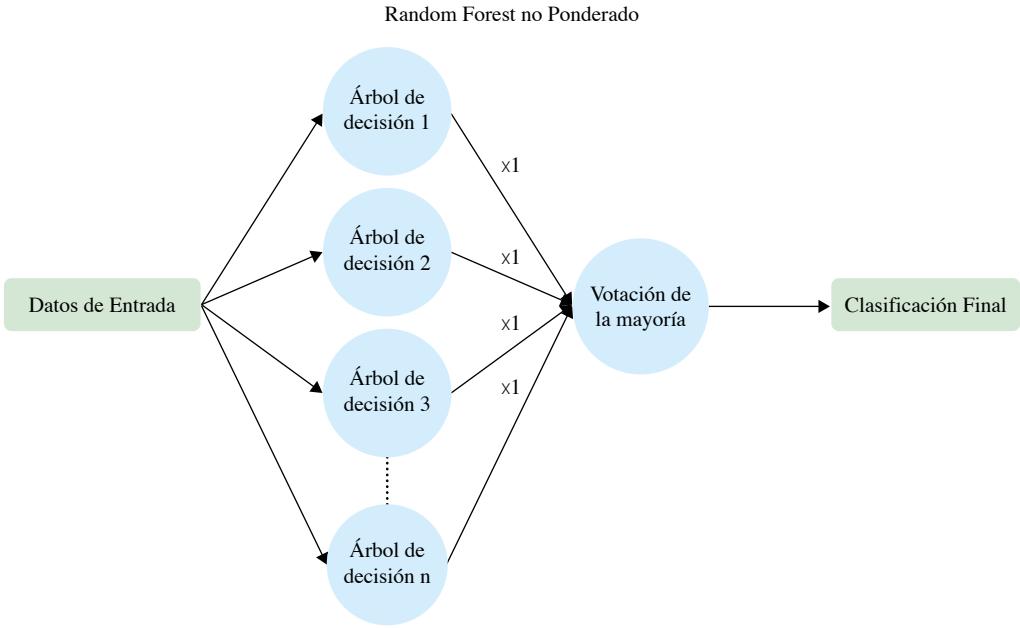
El modelo escogido para desarrollar el algoritmo de predicción de quiebra fue el Random Forest que, según su documentación (RDocumentation, 2021), implementa el algoritmo de Breiman para clasificación y regresión basado en la generación aleatoria de árboles de decisión y una clasificación final basada en mayor cantidad de votos.

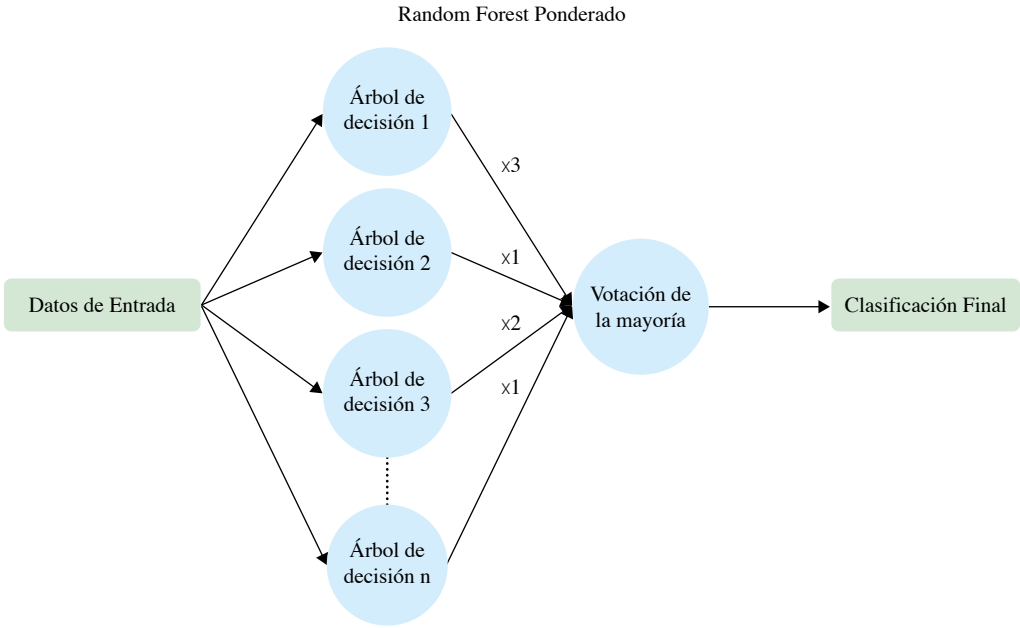
### 3.11.2 El modelo Random Forest en R

La definición más sencilla del algoritmo de Random Forest es: clasificación de un set de datos escogiendo la mayor votación de una gran cantidad de árboles de decisión creados de forma aleatoria, con variables y valores también aleatorios. Según Frederick Livingston, en su documento "Implementation of Breiman's Random Forest Machine Learning", el algoritmo de Random Forest de Breiman se denomina "meta aprendiz", ya que consiste en múltiples aprendices individuales; en este caso, árboles de decisión para realizar un mejor aprendizaje.

Si bien cada árbol se crea de manera aleatoria y por consiguiente el resultado de cada uno es distinto, el algoritmo de Breiman usa un modelo de votaciones en la que se escoge la clasificación más votada por todos los árboles de decisión (Livingston, 2005). Los siguientes gráficos muestran una representación visual del algoritmo de Random Forest no ponderado y ponderado, en los que la diferencia es que se le otorga más peso a los votos de árboles que han demostrado un mayor acierto al momento de clasificar.

Figura 7. Ejemplo de Random Forest Ponderado y no Ponderado






Fuente: Trappenberg (2020).

Se usó la función en R Random Forest con la cual se crearon 100 árboles de decisión con el algoritmo CART trabajando con una muestra de 586 registros Estables y 584 registros de Quiebra; los registros restantes los usa el algoritmo para ejecutar la validación y mostrar los resultados de reducción de error e importancia de las variables. Si se quisiera escoger un solo árbol del bosque se puede usar la función *getTree(rfobj, k=1, labelVar=FALSE)* donde k es el número de árbol a escoger.

Dado que el presente trabajo solo tiene variables numéricas, cada rama del árbol de decisión se crea con base en un valor de corte por cada variable, en donde a la izquierda del árbol quedan aquellas menores o iguales que ese valor y a la derecha las restantes. El valor donde se divide la rama se encuentra reduciendo la impureza de Gini, que no es otra cosa que buscar el valor donde se pueden clasificar todos los registros de una categoría a la derecha y las demás a la izquierda. En otras palabras, reducir la probabilidad de que un elemento elegido aleatoriamente del conjunto sea etiquetado incorrectamente.

Adicionalmente, el algoritmo entrega los resultados de importancia de las variables en el momento de clasificar, en este caso, una empresa en quiebra o estable. La importancia de las variables permite determinar cuáles indicadores

financieros son los que tienen mayor poder para predecir una quiebra y cuáles se podrían predecir para lograr un modelo generalizado (Livingston, 2005). La importancia de la variable es calculada con base en permutaciones aleatorias en diferentes conjuntos de datos en los que se evalúa qué tanto esta variable afectó la clasificación correcta de los registros. A mayor grado de clasificación correcta, mayor es la importancia de la variable.

### 3.12 Resultados del set de entrenamiento y validación

Se presentan los primeros registros del set de entrenamiento, el cual contiene tanto el *LABEL* o marca Quiebra/Estable, así como las variables necesarias para entrenar y validar los tres modelos propuestos. Para cada modelo se seleccionarán únicamente las variables que le corresponden dentro de los parámetros de la función Random Forest, que serán detalladas a continuación.

Figura 8. Visualización del set de entrenamiento

| | LABEL | CapTrabActTot | UtRetActTot | UTOpToAct | VentasTotalesActivos | MargenOp | ROE | CrecimientoVentas | CrecimientoActivo | ActPAs | PercPasCP | MargNeto | RazonCorriente |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Quiebra | 0.049031817 | 0.025106128 | 0.0060725161 | 0.62564146 | 0.020997018 | 1.407068e-02 | 0.248899096 | 0.3875891507 | 1.7592387 | 1.000000000 | 9.706064e-03 | 1.08625867 |
| 2 | Quiebra | 0.049031817 | 0.025106128 | 0.0060725161 | 0.62564146 | 0.020997018 | 1.407068e-02 | 0.248899096 | 0.3875891507 | 1.7592387 | 1.000000000 | 9.706064e-03 | 1.08625867 |
| 3 | Quiebra | 0.049031817 | 0.025106128 | 0.0060725161 | 0.62564146 | 0.020997018 | 1.407068e-02 | 0.248899096 | 0.3875891507 | 1.7592387 | 1.000000000 | 9.706064e-03 | 1.08625867 |
| 4 | Quiebra | 0.104037688 | 0.031138322 | 0.0036989748 | 0.69511885 | 0.025666122 | 1.143525e-02 | -0.124746536 | 0.0384866885 | 1.4781339 | 1.000000000 | 5.321356e-03 | 1.15378163 |
| 5 | Quiebra | 0.104037688 | 0.031138322 | 0.0036989748 | 0.69511885 | 0.025666122 | 1.143525e-02 | -0.124746536 | 0.0384866885 | 1.4781339 | 1.000000000 | 5.321356e-03 | 1.15378163 |
| 6 | Quiebra | 0.104037688 | 0.031138322 | 0.0036989748 | 0.69511885 | 0.025666122 | 1.143525e-02 | -0.124746536 | 0.0384866885 | 1.4781339 | 1.000000000 | 5.321356e-03 | 1.15378163 |
| 7 | Quiebra | 0.088335416 | 0.040501374 | -0.1034854910 | 0.72964748 | -0.072463757 | -2.715958e-01 | -0.012124368 | 0.1139765845 | 1.6155810 | 0.538527830 | -1.418294e-01 | 1.26500584 |
| 8 | Quiebra | 0.088335416 | 0.040501374 | -0.1034854910 | 0.72964748 | -0.072463757 | -2.715958e-01 | -0.012124368 | 0.1139765845 | 1.6155810 | 0.538527830 | -1.418294e-01 | 1.26500584 |
| 9 | Quiebra | 0.088335416 | 0.040501374 | -0.1034854910 | 0.72964748 | -0.072463757 | -2.715958e-01 | -0.012124368 | 0.1139765845 | 1.6155810 | 0.538527830 | -1.418294e-01 | 1.26500584 |
| 10 | Quiebra | 0.030534402 | 0.086517467 | 0.0033141163 | 1.02469930 | 0.064207808 | 1.846247e-02 | -0.301101790 | 0.5346692218 | 1.2187774 | 0.996498321 | 3.234233e-03 | 1.03727061 |
| 11 | Quiebra | 0.030534402 | 0.086517467 | 0.0033141163 | 1.02469930 | 0.064207808 | 1.846247e-02 | -0.301101790 | 0.5346692218 | 1.2187774 | 0.996498321 | 3.234233e-03 | 1.03727061 |
| 12 | Quiebra | 0.030534402 | 0.086517467 | 0.0033141163 | 1.02469930 | 0.064207808 | 1.846247e-02 | -0.301101790 | 0.5346692218 | 1.2187774 | 0.996498321 | 3.234233e-03 | 1.03727061 |
| 13 | Quiebra | 0.661031895 | -0.797725692 | -0.0036915601 | 1.54853790 | -0.003965504 | -2.439718e-03 | -0.089584581 | -0.0860500025 | 0.3979135 | 0.070729734 | -2.383900e-03 | 4.71885326 |
| 14 | Quiebra | 0.661031895 | -0.797725692 | -0.0036915601 | 1.54853790 | -0.003965504 | -2.439718e-03 | -0.089584581 | -0.0860500025 | 0.3979135 | 0.070729734 | -2.383900e-03 | 4.71885326 |
| 15 | Quiebra | 0.661031895 | -0.797725692 | -0.0036915601 | 1.54853790 | -0.003965504 | -2.439718e-03 | -0.089584581 | -0.0860500025 | 0.3979135 | 0.070729734 | -2.383900e-03 | 4.71885326 |
| 16 | Quiebra | 0.695536939 | -0.757371774 | 0.0282898792 | 1.55454991 | 0.013619364 | -2.050686e-02 | -0.114622669 | 0.0010892768 | 0.4202506 | 0.103017127 | 1.819812e-02 | 3.83739037 |
| 17 | Quiebra | 0.695536939 | -0.757371774 | 0.0282898792 | 1.55454991 | 0.013619364 | -2.050686e-02 | -0.114622669 | 0.0010892768 | 0.4202506 | 0.103017127 | 1.819812e-02 | 3.83739037 |
| 18 | Quiebra | 0.695536939 | -0.757371774 | 0.0282898792 | 1.55454991 | 0.013619364 | -2.050686e-02 | -0.114622669 | 0.0010892768 | 0.4202506 | 0.103017127 | 1.819812e-02 | 3.83739037 |
| 19 | Quiebra | -0.206198434 | -0.710028184 | -0.0118345177 | 0.00969494 | -0.942604528 | -7.047952e-02 | -0.314366413 | -0.0039714475 | 1.2017992 | 0.272263804 | -1.220690e+00 | 0.08981981 |
| 20 | Quiebra | -0.206198434 | -0.710028184 | -0.0118345177 | 0.00969494 | -0.942604528 | -7.047952e-02 | -0.314366413 | -0.0039714475 | 1.2017992 | 0.272263804 | -1.220690e+00 | 0.08981981 |
| 21 | Quiebra | -0.206198434 | -0.710028184 | -0.0118345177 | 0.00969494 | -0.942604528 | -7.047952e-02 | -0.314366413 | -0.0039714475 | 1.2017992 | 0.272263804 | -1.220690e+00 | 0.08981981 |
| 22 | Quiebra | 0.462731119 | 0.053005613 | -0.0008605997 | 0.66788763 | 0.044648479 | -2.459359e-03 | -0.348364930 | 0.1429395213 | 1.5382922 | 0.438410477 | -1.288540e-03 | 2.62362830 |
| 23 | Quiebra | 0.462731119 | 0.053005613 | -0.0008605997 | 0.66788763 | 0.044648479 | -2.459359e-03 | -0.348364930 | 0.1429395213 | 1.5382922 | 0.438410477 | -1.288540e-03 | 2.62362830 |
| 24 | Quiebra | 0.462731119 | 0.053005613 | -0.0008605997 | 0.66788763 | 0.044648479 | -2.459359e-03 | -0.348364930 | 0.1429395213 | 1.5382922 | 0.438410477 | -1.288540e-03 | 2.62362830 |
| 25 | Quiebra | -0.149670027 | -0.228725639 | -0.0130720890 | 0.49385516 | -0.020379815 | -2.879064e-02 | -0.119859361 | -0.0055528886 | 1.8316346 | 0.793355890 | -2.646948e-02 | 0.65445419 |
| 26 | Quiebra | -0.149670027 | -0.228725639 | -0.0130720890 | 0.49385516 | -0.020379815 | -2.879064e-02 | -0.119859361 | -0.0055528886 | 1.8316346 | 0.793355890 | -2.646948e-02 | 0.65445419 |
| 27 | Quiebra | -0.149670027 | -0.228725639 | -0.0130720890 | 0.49385516 | -0.020379815 | -2.879064e-02 | -0.119859361 | -0.0055528886 | 1.8316346 | 0.793355890 | -2.646948e-02 | 0.65445419 |
| 28 | Quiebra | -0.211384108 | -1.571139618 | -0.0503103839 | 0.48086343 | -0.484796747 | -5.470810e-02 | -0.547298181 | -0.2101593577 | 0.5209378 | 0.183978104 | -1.046251e-01 | 0.40146152 |

Fuente: elaboración propia.

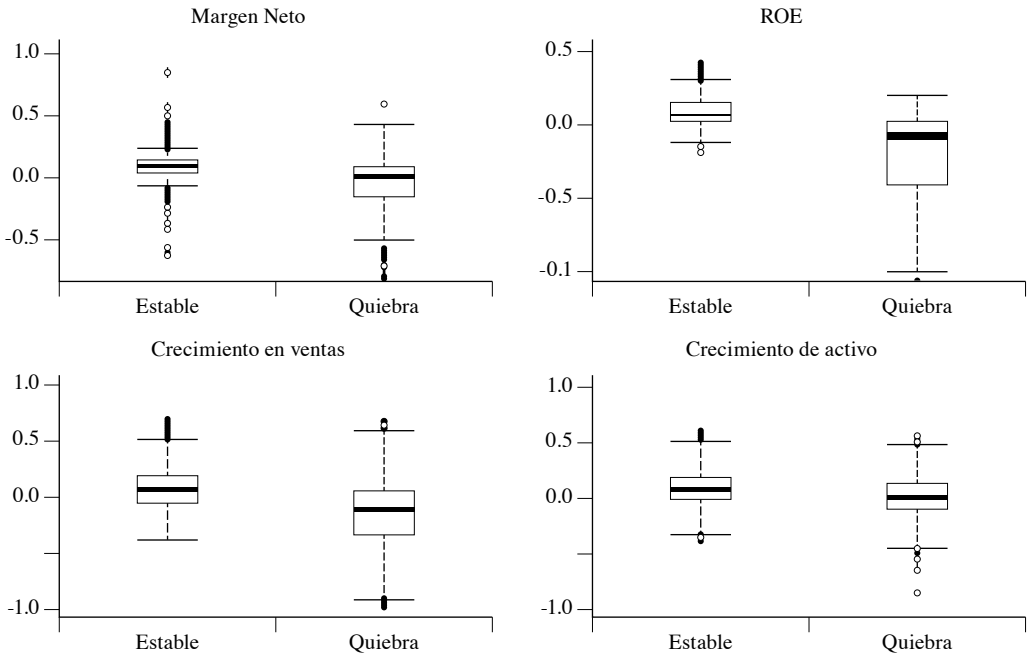
### 3.12.1 Resultados de entrenamiento y validación Altman Z-Score ajustado

Para el Z-Score de empresas colombianas se crearon 100 árboles de decisión, con una muestra de 586 datos estables y 584 datos de quiebra y 2 variables por iteración.

La fórmula aplicada en R para el entrenamiento:

*randomForest(formula = label ~ CapTrabActTot + UtRetActTot + UTNetaToAct + VentasTotalesActivos, data = the.data, ntree = 100, replace = TRUE, sampsize = c(586, 584))*

Figura 9. SEQ Figura * ARABIC \s 1 9 *Box-Plot* Variables Altman Z-

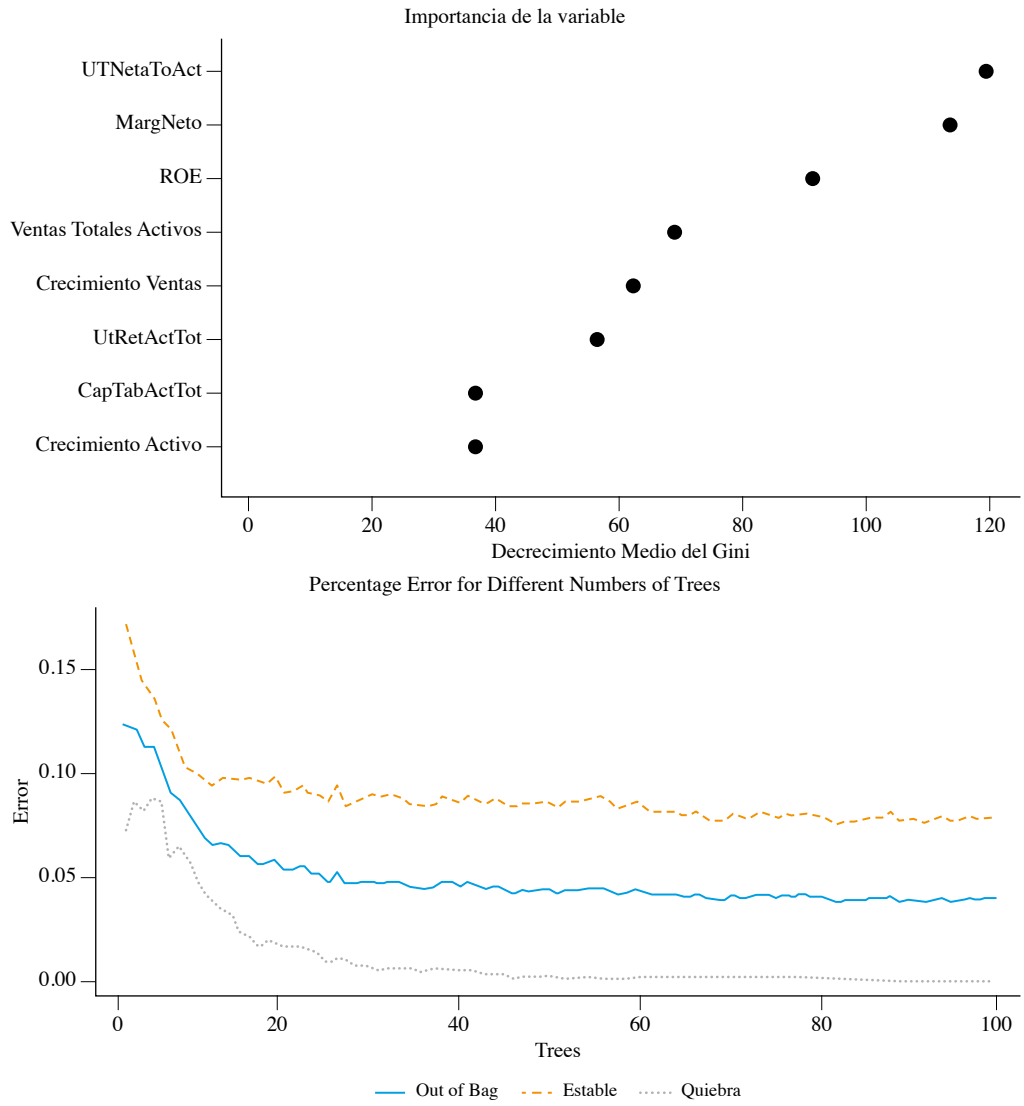



Fuente: elaboración propia.

Para el modelo Z-Score de Altman, ajustado a 4 variables, se puede evidenciar en el *box plot* que las medianas y las distribuciones de los datos para las variables de ROA y Ventas Totales/Activos Totales presentan la mayor diferencia entre las empresas que estarán estables en 1 o 2 años, frente a las que entrarán en quiebra en el mismo periodo.

El algoritmo Random Forest lo confirma con el gráfico de decrecimiento de Gini en donde UTNetaToAct es la variable que más aporta en el momento de generar segmentaciones en las ramas del árbol. Adicionalmente, con 100 árboles se lograron los niveles más bajos tanto en error de Quiebra como en error de Estable.

Figura 10. SEQ Figura * ARABIC \s 1 10 Resumen de Entrenamiento Altman Z-Score ajustado




Fuente: elaboración propia.

Tabla 8. Resultados de v*alidación* Altman Z-Score Ajustado

| Modelo | Verdaderos Positivos | Verdaderos Negativos | Falsos Positivos | Falsos Negativos | % Error Tipo I | % Error Tipo II | AUC |
|---|---|---|---|---|---|---|---|
| Altman Z-Score Colombia | 834 | 1124 | 47 | 333 | 28,7 | 4,01 | 93,3 |

Fuente: elaboración propia.

La matriz de confusión en la validación mostró un error del 4% en la predicción de empresas estables y 28,7% en la predicción de quiebras. Las métricas anteriores muestran una reducción en la capacidad predictiva mostrado por la validación automática que realiza el algoritmo CART en la que el error Tipo I descendió al 11% y el Error Tipo II al 0%. Estos resultados eran de esperarse, ya que para balancear el set de datos se duplicaron registros con marca quiebra que llegan a repetirse a medida que se crea una mayor cantidad de árboles de decisión y el algoritmo tiende a reconocer datos ya vistos. En este caso, se permite este sobreajuste de la clasificación Quiebra y, por consiguiente, una mayor cantidad de falsos positivos Quiebra, ya que se busca crear una alerta temprana para tomar medidas correctivas. Es preferible alertar de una quiebra que podría no llegar a suceder, antes que marcar como estable una empresa que puede irse a la quiebra.

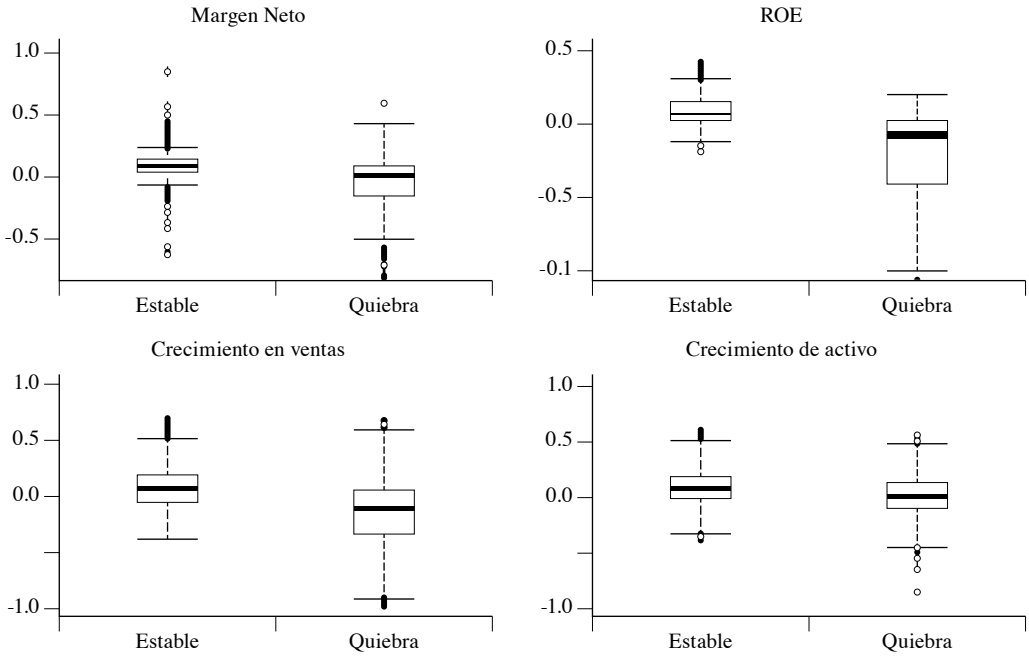
### 3.12.2 Resultados de entrenamiento y validación–Altman, Barboza y Kimura ajustado

Para el modelo Barboza *et al*. (2017), ajustado a empresas colombianas, se crearon 100 árboles de decisión con una muestra de 586 datos estables y 584 datos de quiebra, 2 variables usadas en cada iteración.

La fórmula aplicada en R para el entrenamiento:

*randomForest(formula = LABEL ~ CapTrabActTot + UtRetActTot + UTNetaToAct + VentasTotalesActivos + ROE + CrecimientoVentas + CrecimientoActivo + MargNeto s, data = the.data, ntree = 100, replace = TRUE, sampsize = c(586, 584))*

Figura 11. STYLEREF 1 \s 6. SEQ Figura * ARABIC \s 1 11 Box-Plot Variables Altman, BarbozaKimura

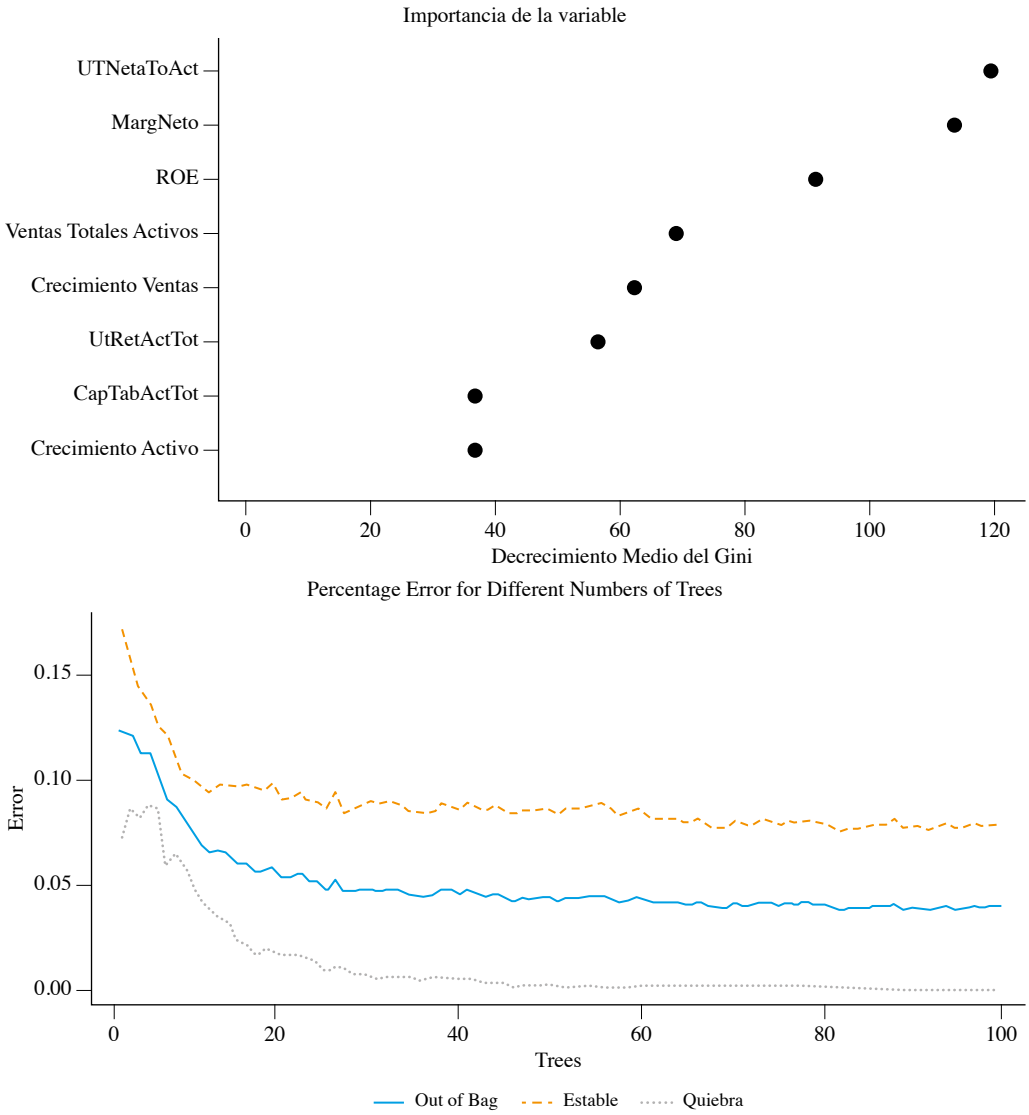



Fuente: elaboración propia.

La exploración gráfica por *box-plot* muestra que las medianas y los rangos intercuartiles del margen neto y el ROE tienden a diferenciarse, lo que el algoritmo confirma posicionándolas en el segundo y tercer puesto de importancia de la variable. El crecimiento del activo se queda en el último lugar ante la similitud en las distribuciones entre las dos marcas.

De nuevo, con 100 árboles se lograron los niveles más bajos tanto en error de Quiebra como en error de Estable, pero con una disminución mayor del error de clasificación de empresas estables. De nuevo se llevó el error de Quiebra a 0% ante el entrenamiento exhaustivo de esta categoría. Se mantiene la metodología mencionada en el primer modelo, en búsqueda de una alerta de quiebra temprana.

Figura 12. STYLEREF 1 \s 6. SEQ Figura * ARABIC \s 1 12 Resumen de entrenamiento Altman, Barboza Kimura ajustado




Fuente: elaboración propia.

Tabla 9. Resultados de validación Altman, Barboza y Kimura ajustado

| Modelo | Verdaderos Positivos | Verdaderos Negativos | Falsos Positivos | Falsos Negativos | % Error Tipo I | % Error Tipo II | AUC |
|---|---|---|---|---|---|---|---|
| Altman, Barboza, Kimura Colombia | 919 | 1139 | 32 | 248 | 21,2 | 2,7 | 94,5 |

Fuente: elaboración propia.

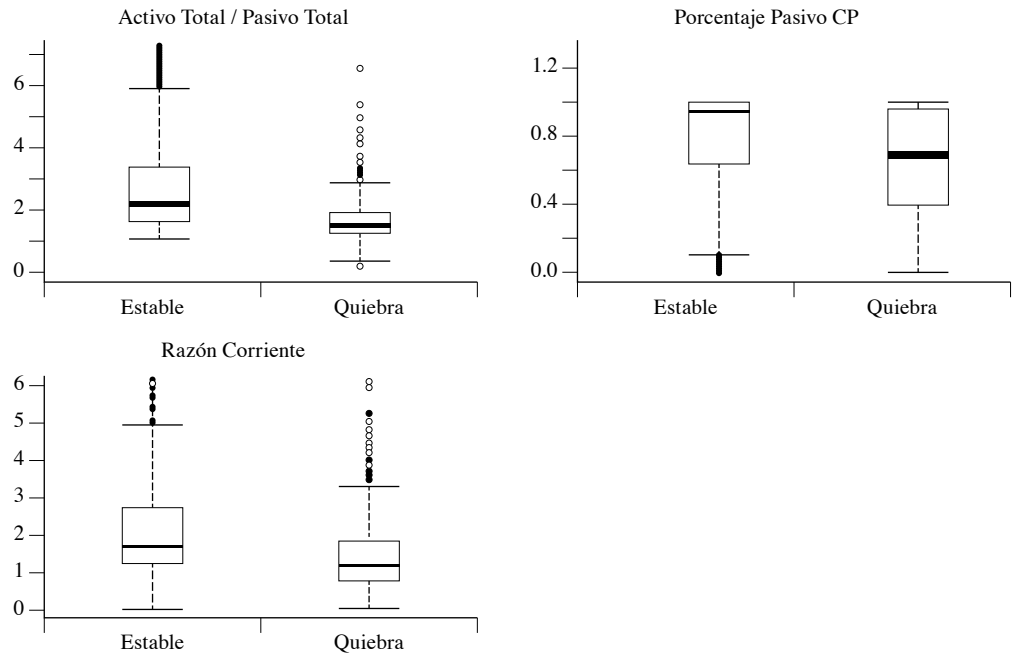
### 3.12.3 Resultados de entrenamiento y validación del modelo propuesto

En el modelo propuesto para las empresas colombianas se crearán 100 árboles de decisión, con una muestra de 586 datos Estables y 584 datos de Quiebra, 2 variables usadas en cada iteración.

La fórmula aplicada en R para el entrenamiento:

*randomForest(formula = LABEL ~ CapTrabActTot + UtRetActTot + UTOpToAct + VentasTotalesActivos + ROE + CrecimientoVentas + CrecimientoActivo + ActPAs + PercPasCP + MargNeto + RazonCorriente, data = the.data, ntree = 100, replace = TRUE, sampsize = c(586, 584))*

Figura 13. STYLEREF 1 \s 6. SEQ Figura * ARABIC \s 1 13 -plot modelo propuesto

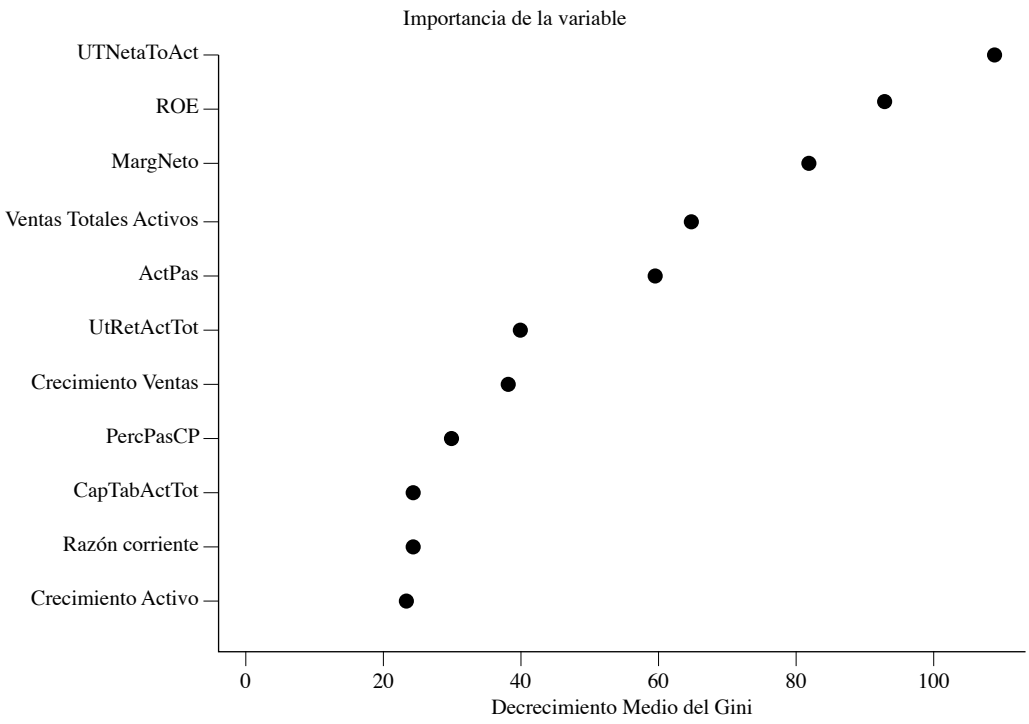


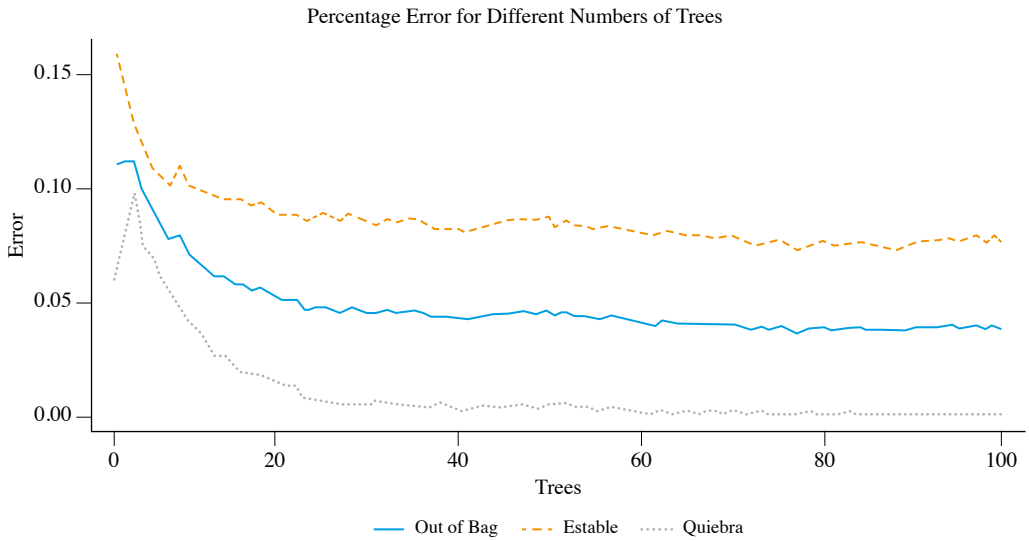

Fuente: elaboración propia.

La exploración gráfica por *box-plot* muestra que las medianas de las tres nuevas variables son visualmente diferentes, sin embargo, los rangos intercuartiles tienden a sobreponerse con una importante presencia de datos atípicos, tanto en el Activo Total / Pasivo Total como en la razón corriente. Estas características agregaron valor al modelo, pero las nuevas variables no superaron en importancia las escogidas en los modelos anteriores. Pese a esto, se logró reducir el error de clasificación de empresas estables de 7,9 a 7,6%.

Aunque con 100 árboles se lograron los niveles más bajos de error, la estabilidad de este se da a partir de los 83 árboles, por tanto, se puede reducir la cantidad de árboles de decisión a generar a este número. De nuevo, se llevó el error de Quiebra a 0% ante el entrenamiento exhaustivo de esta categoría. Se mantiene la metodología mencionada en el primer modelo en búsqueda de una alerta de quiebra temprana siendo aceptable el sobreajuste y la tasa de error en la categoría Quiebra. Se quiere ser ácido y alertar quiebras con una mayor probabilidad de acierto, pese a que se penalicen empresas que podrán esquivar este estado.

Figura 14. Resumen de resultados entrenamiento modelo






Fuente: elaboración propia.

Tabla 10. Resultados de validación en el modelo propuesto

| Modelo | Verdaderos Positivos | Verdaderos Negativos | Falsos Positivos | Falsos Negativos | % Error Tipo I | % Error Tipo II | AUC |
|---|---|---|---|---|---|---|---|
| Propuesta Colombia | 950 | 1.118 | 53 | 217 | 18,6 | 4,5 | 95,7 |

Fuente: elaboración propia.

## 3.13 Resumen de entrenamiento y validación

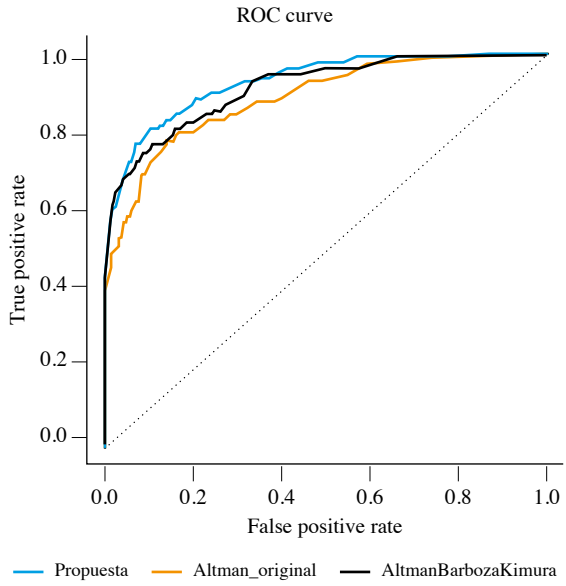
- Con los datos de 1995 hasta 2015 se entrenaron y validaron los modelos. Estos datos fueron limpiados y balanceados para enviarle la mayor cantidad de información al algoritmo de las características de una empresa en quiebra.
- Se destinaron los datos de 2016 para evaluar cada modelo. Estos datos no fueron balanceados para establecer el comportamiento en un escenario real del algoritmo.
- Se entrenaron los modelos con sus correspondientes variables haciendo uso del algoritmo de Random Forest con 100 árboles. Se permitió el sobreajuste al duplicar registros de quiebra, lo que llevó la tasa de error de la marca Quiebra a 0% en la validación dentro del modelo CART. Sin embargo, el set de validación mostró resultados favorables que más adelante, al ser comparados con el set de evaluación descartan el sobreajuste del modelo.

- Se permite una mayor tasa de error al clasificar empresas que fueron estables como Quiebra con el fin de permitirle al algoritmo ser un indicador ácido de alerta temprana a 2 años.
- Se evaluaron los resultados con un set completamente desconocido de datos de 2016 y se calcularon las métricas de evaluación para determinar la combinación de variables que ofrece el mayor poder predictivo.

### 3.14 Resultados del set de evaluación

Con los modelos entrenados se realizó la clasificación del set de evaluación. Lo relevante de este set es que son datos que el algoritmo no ha visto y, por lo tanto, son la referencia clave para determinar el éxito y la aplicación del modelo. No se penalizará la diferencia entre los errores tipo I y tipo II entre el set de validación y el set de evaluación; lo anterior debido a la duplicidad de registros para balancear el set de entrenamiento. La evaluación general se realizará con base en el área bajo la curva ROC (AUC).

Figura 15. SEQ figura * ARABIC \s 1 12 curva ROC - Comparación




Fuente: elaboración propia.

Como se presenta en la figura 15 y en la tabla 11, a medida que se incluyeron indicadores financieros, los modelos ganaron en poder predictivo, siendo la propuesta de este trabajo la mayor área bajo la curva.

Al revisar los demás indicadores, el modelo propuesto queda en segundo lugar en cuanto a F1 Score y Precisión, debido al sacrificio entre predicción de empresas en quiebra a costa de una reducción en la predicción de empresas estables. Sin embargo, dado que el énfasis está en la predicción de quiebras y la intención de una predicción más ácida para dar una alerta más efectiva, la ganancia en AUC justifica la preferencia del modelo propuesto frente a los demás.

Tabla 11. Resultados del set de evaluación

| Modelo | Precisión % | F1 Score | AUC % | Precisión Estable % | Precisión Quiebra % |
|---|---|---|---|---|---|
| Altman Z- Score | 93,25 | 0,9646 | 88,89 | 93,89 | 60,31 |
| Altman, Barboza, Kimura | 94,95 % | 0,9738 | 91,30 | 95,55 | 68,70 |
| Propuesta | 93,33 | 0,9649 | 92,89 | 93,76 | 74,05 |

Fuente: elaboración propia.

Por último, si se comparan los resultados con los del trabajo original de Barboza *et al*. (2017), se evidencia que la metodología usada en los tres modelos permitió reducir considerablemente el error de clasificar una empresa que entrará en quiebra como Estable, a costa de clasificar como en posible Quiebra a las empresas que estarán estables en dos años.

Este intercambio de clasificación y un mayor error en las empresas en quiebra va alienado con el fin de establecer una alerta temprana a dos años. Es importante aclarar que, a mayor tiempo hasta la quiebra, se pueden tomar mejores medidas, por lo que la probabilidad de evitarla es más alta y, por lo tanto, también la posibilidad de que el algoritmo falle. Por lo anterior, se puede afirmar que el modelo propuesto sirve de alerta a los administradores de la empresa para tomar de manera anticipada decisiones que permitan mantener la estabilidad de la compañía en el mediano plazo.

Tabla 12. Comparativo de evaluación de los modelos

| Modelo | Verdaderos Positivos | Verdaderos Negativos | Falsos Positivos | Falsos Negativos | % Error Tipo I | % Error Tipo II | AUC |
|---|---|---|---|---|---|---|---|
| Random Forest Altman, Barboza, Kimura | 112 | 12.536 | 2017 | 35 | 23,81 | 13,86 | 91,15 |

| Modelo | Verdaderos Positivos | Verdaderos Negativos | Falsos Positivos | Falsos Negativos | % Error Tipo I | % Error Tipo II | AUC |
|---|---|---|---|---|---|---|---|
| Altman Z-Score Colombia | 79 | 5531 | 352 | 52 | 39,7 | 6 | 88,8 |
| Altman, Barboza, Kimura Colombia | 90 | 5623 | 262 | 41 | 31,3 | 4,5 | 91,3 |
| Propuesta Colombia | 97 | 5518 | 367 | 34 | 26 | 6,2 | 92,9 |

Fuente: elaboración propia con base en Rosillo (2002); Trappenberg (2020).

En cuanto al desempeño del modelo frente al set de validación y de evaluación, se experimentó un descenso en el poder predictivo de la categoría Quiebra cercana al 10% y 2% para la categoría Estable. El AUC se redujo en promedio 4% por lo que no se evidencia sobreajuste del modelo, el cual cumple con la función de alertar una posible quiebra al permitirse un mayor error tipo I que tipo II.

## Conclusiones

En conclusión, tanto el modelo de Altmn, Barboza y Kimura como el modelo propuesto en este estudio han demostrado ser herramientas confiables y útiles para alertar sobre una posible quiebra empresarial a dos años. La inclusión de nuevas variables en el modelo propuesto incrementó el área bajo la curva ROC (AUC), lo que refuerza su capacidad como una alerta temprana para tomar decisiones preventivas que ayuden a evitar la quiebra.

En el contexto de la predicción de quiebra, es preferible diagnosticar una compañía sana en riesgo de quiebra que fallar en identificar a una empresa en riesgo real como sana. El ajuste en el entrenamiento, buscando evitar el sobreajuste, fue crucial para balancear los resultados hacia este objetivo. La estrategia permitió optimizar el modelo y mejoró la identificación de empresas con alta probabilidad de quiebra.

Para futuros trabajos, se propone que, con mejor información proporcionada por entidades públicas colombianas, se incluyan métricas relacionadas con el flujo de caja y el EBITDA, así como indicadores macroeconómicos líderes que permitan aumentar el poder predictivo e, incluso, anticiparse no solo a dos, sino hasta tres años. Además, se recomienda excluir los años 2020 hasta posiblemente 2022 en los entrenamientos futuros, ya que la pandemia del covid-19 hizo que estos años se comportaran de manera atípica, lo que no permite

generar un modelo general. Estos años deben reservarse exclusivamente para la evaluación de escenarios de estrés.

## Referencias

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589-609. https://doi.org/10.2307/2978933

Barboza, F., Kimura, H. y Altman, E. (2017). *Machine Learning models and Bankrupcy prediction*.

Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research, 4*, 71-111.

Decreto 2784 de 2012. *Marco técnico normativo para los preparadores de información financiera que conforman el Grupo 1 en* https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=75511

Livingston, F. (2005). *ECE591Q Machine Learning journal paper: Implementation of Breiman's Random Forest Machine Learning algorithm*.

Ramírez Luna, S., Roa, E. M., Ariza, M. y Ferrín, H. (2015). *Modelos de predicción de alerta temprana para riesgos de quiebra de pymes sector industrial de Bogotá*.

Rdocumentation. (2021). *RandomForest: Classification and regression with Random Forest*. https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest

*Reportes de información f*inanciera *para empresas del grupo 1 y grupo 2 de la Superintendencia de Sociedades de 1995 a 2016 en* https://www.supersociedades.gov.co/tramites-opa-y-consulta-de-informacion/-/asset_publisher/ltml/content/siis-consulta-de-estados-financieros-hist%25C3%25B3ricos-1.

Romero Espinosa, F. (2013). Alcances y limitaciones de los modelos de capacidad predictiva en el análisis del fracaso empresarial. AD-minister, (23), 45-70.

Rosillo, J. (2002). Modelo de predicción de quiebras de las empresas colombianas. *INNOVAR, Revista de ciencias administrativas y sociales*, *19*, 109-124.

Tascón Fernández, M. T. y Castaño Gutiérrez, F. J. (2012). *Variables y modelos para la identificación y predicción del fracaso empresarial: Revisión de la investigación empírica reciente*. S. D.

Trappenberg, T. P. (2020). *Fundamentals of Machine Learning*. Dalhousie University, Oxford University Press.