COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad Ciencias Económicas y Empresariales

PREDICCIÓN DE QUIEBRAS EN EL SECTOR DE LA RESTAURACIÓN EN ESPAÑA

Autor: Lucía Morrás Lorenzo

Director: María Lourdes Fernández Rodríguez

MADRID | Abril 2023

RESUMEN

En este trabajo se trata de predecir la quiebra de las empresas recogidas en el sector restaurantes y puestos de comida según la CNAE-2009, dentro del sector de la restauración en España. El objeto del análisis es predecir la quiebra con un año de antelación, con el propósito de servir de inspiración a empresas dedicadas a las fusiones y adquisiciones a implementar métodos novedosos para ayudar a tomar decisiones en torno a qué entidades adquirir. Para lograr este objetivo, se utilizan dos métodos de modelado: la regresión logística y las redes neuronales, implementadas en el entorno Spyder de Python. Los resultados muestran que los modelos pueden predecir la quiebra con un año de antelación, pero aún tienen margen de mejora. En cuanto a la información empleada para el análisis, el estudio se basa en información contable del último año disponible en la base de datos SABI, por lo que sería interesante considerar la incorporación de información no contable en los modelos para comprobar su impacto en la precisión predictiva.

PALABRAS CLAVE

Predicción de quiebras; regresión logística; redes neuronales; información contable; ratios financieros

ABSTRACT

The aim of this paper is to predict the bankruptcy of the companies included in the restaurant and food stalls sector according to CNAE-2009, within the catering sector in Spain. The aim of the analysis is to predict bankruptcy one year in advance, with the purpose of inspiring companies involved in mergers and acquisitions to implement novel methods to help them make decisions about which entities to acquire. To achieve this goal, two modelling methods are used: logistic regression and neural networks, implemented in Python's Spyder environment. The results show that the models can predict bankruptcy one year in advance, but there is still room for improvement. As for the information used for the analysis, the study is based on accounting information for the last year available in the SABI database, so it would be interesting to consider incorporating non-accounting information in the models to check its impact on predictive accuracy.

KEY WORKS

Bankruptcy prediction; logistic regression; neural networks; accounting information; financial ratios

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO 1. INTRODUCCIÓN

1.1. MOTIVACIÓN

La idea para llevar a cabo este trabajo fin de grado se inspira en la relevancia que tiene el sector de la restauración en España, siendo este un elemento de cohesión social en la cultura española. En este sentido, en el contexto de crisis económica actual que sigue tratando de paliar los efectos de la pandemia, la quiebra empresarial ha suscitado un especial interés en mí. Todo ello, conjugado con la creciente consolidación de grupos de restauración ha confirmado la necesidad de un estudio relativo a la quiebra en el sector empresarial en el marco de las fusiones y adquisiciones.

La manera elegida para consolidar un estudio que reúna el análisis de quiebra empresarial y el contexto económico actual ha venido dada por la culminación de mis estudios en *Business Analytics*, y no podía ser otra forma que mediante la implementación de modelos predictivos.

1.2. OBJETIVO

El objetivo de este trabajo es proporcionar a las empresas un estudio empírico de la implementación de modelos de predicción de quiebras que arrojen información y técnicas relevantes en el marco de las fusiones y adquisiciones. En esta línea, teniendo en cuenta la creciente consolidación de grupos de restauración, se trata de, con las técnicas implementadas, proporcionar métodos para obtener una lista de empresas que sería proclive a la quiebra, aunque de *facto* no lo hayan hecho, para que las entidades adquirentes puedan observar la evolución de las compañías y cuenten con un poder de negociación añadido.

1.3. METODOLOGÍA

La forma de proceder para desarrollar este trabajo implica iteración en las ejecuciones de los modelos modificando sus características. Por ello, para la implementación de las técnicas de *machine learning* elegidas, se han ido creando diferentes escenarios que tratan de mejorar los resultados obtenidos en primer lugar. El proceso seguido va desde la extracción de datos, hasta la obtención de resultados, pasando por la creación y análisis de variables. Se explicarán todos los conceptos de la ciencia de datos utilizados en su correcto contexto para facilitar la comprensión del estudio.

Es importante destacar la importancia de contar con una base de datos fiable para poder desarrollar el análisis que pretende el trabajo. En este caso, se ha obtenido la información disponible en la base de datos SABI (por sus siglas, Sistema de Análisis de Balances Ibéricos).

Esta base de datos cuenta con información de prácticamente cualquier empresa con sede en Iberia, lo que facilita la acotación del trabajo y asegura que la información es fidedigna.

1.4. CÓDIGO

La plataforma elegida para desarrollar este trabajo ha sido Spyder, dentro del entorno de Python. Esta plataforma permite de manera dinámica realizar diferentes pruebas de manera sencilla. Sin embargo, el motivo principal de haber elegido Spyder por encima del resto de entornos es su habitual uso en el ámbito empresarial. De esta forma, como el objetivo de este trabajo fin de grado es el de dotar de herramientas novedosas a las empresas para sus análisis, se ha tratado de usar una plataforma que con la que ya estén familiarizadas.

El código, que se encuentra en el Anexo, cuenta con los comentarios necesarios para poder seguir la implementación de los diferentes escenarios y técnicas utilizadas paso a paso, aún sin estar familiarizado con la programación.

CAPÍTULO 2. PREDICCIÓN DE QUIEBRAS

2.1. CONCEPTO DE QUIEBRA

Según la Real Academia Española (2023), la quiebra es el "juicio por el que se incapacita patrimonialmente a alguien por su situación de insolvencia y se procede a ejecutar todos sus bienes en favor de la totalidad de sus acreedores". Trasladada la definición al entorno empresarial, la quiebra hace referencia a la situación en la que una empresa no puede hacer frente a sus obligaciones financieras, es decir no puede pagar sus deudas.

Desde el punto de vista contable, la quiebra es la situación patrimonial en la que el valor del activo registrado en las partidas contables es inferior al valor del pasivo exigible. La situación de quiebra es fácilmente visible en el balance de situación de una empresa, y se dará cuando el ratio de solvencia[1] sea menor a 1.

Desde la perspectiva legal, es preciso mencionar diferentes conceptos técnicos que aluden a problemas económicos, como son la suspensión de pagos, la quiebra y el concurso de acreedores, todos ellos relacionados con la imposibilidad de una empresa de satisfacer sus deudas. La suspensión de pagos sería el primer supuesto, también llamada "quiebra técnica"; se caracteriza por ser reversible pues, tan solo hace alusión a la falta de liquidez en sí misma. En este caso la empresa tiene activos suficientes para afrontar las deudas, pero no liquidez monetaria. En la definición de "quiebra legal" la empresa deudora no posee activo suficiente -de ningún tipo- para afrontar sus deudas, por lo que la situación de falta de liquidez es permanente. En última instancia, el concurso de acreedores o procedimiento de insolvencia es "un procedimiento colectivo sujeto a una supervisión judicial que se sustancia con miras a la reorganización o a la liquidación de una empresa insolvente" (Comisión de las Naciones Unidas para el Derecho Mercantil Internacional, 2006, pág. 6). Desde el punto de vista de la empresa deudora, se trata de un mecanismo legal que trata de resolver la situación de insolvencia, mediante la reorganización o la liquidación con el fin de evitar el cierre definitivo de la empresa. Desde la perspectiva de los acreedores, trata de conseguir que el mayor número de estos perciba la mayor cantidad posible de la deuda exigible, con el fin de saldar la mayor parte de esta.

Es preciso explicar de manera sucinta cómo funcionan los concursos de acreedores en España para ganar un mayor entendimiento del concepto de quiebra. Haciendo referencia al artículo 508 de la Ley Concursal española, existen cuatro posibles fases en los concursos: fase

[1] Total Activo/Total Pasivo

común, fase de convenio, fase de liquidación y fase de calificación, además de la recién añadida fase pre concursal (Jefatura de Estado, 2022).

- La fase común tiene como objetivo la identificación y cuantificación de la deuda que, tras haber nombrado un administrador concursal, pasará a la fase de convenio. En esta fase se trata de conciliar un plan de pagos entre los acreedores y la entidad deudora que permita su viabilidad.
- La fase de liquidación llega en caso de que no se llegue a lograr acuerdo en la fase común, procediendo a la ejecución de los bienes de la empresa, tratando de realizar el activo y pagar a los acreedores.
- Por último, en la fase de calificación se trata de determinar la responsabilidad de los administradores por las deudas.

En España, la tasa de éxito de los concursos de acreedores es menor al 10%, por lo que se puede convenir que una empresa una vez se declara en concurso, en la mayoría de las ocasiones no saldrá adelante, y se estará ante una empresa en quiebra definitiva (Jefatura de Estado, 2022). De esta forma para el propósito de este trabajo se considerará que una empresa "quebrada" incluye tanto las empresas "extinguidas" como así las denomina la base de datos SABI, como a las empresas "en liquidación" (fase concursal), ya que es un prolegómeno del parecer de la empresa, que con casi total seguridad no sobrevivirá.

2.2. PREDICCIÓN DE QUIEBRA: IMPORTANCIA

La predicción de quiebra hace alusión a la capacidad de anticipar si una empresa entrará en situación de quiebra en un futuro a partir de información actual e histórica. En el contexto económico actual español, la alta tasa de mortalidad empresarial es un factor determinante para el estudio de la quiebra, ya que, según un estudio realizado por el Instituto Nacional de Estadística (2022), al menos el 21,4% de las empresas en España no logran sobrevivir después de su primer año de actividad y menos del 50% sobreviven más de 5 años.

En este sentido, la predicción de quiebras es útil para la planificación y gestión de riesgos financieros, pudiendo así las empresas identificar los factores de riesgo más relevantes y los indicadores más claros, y así tomar medidas preventivas para evitarlos. Ser capaces de predecir una quiebra puede suponer evitar pérdidas económicas muy significativas, sobre todo para los grupos de interés, comúnmente denominados *stakeholders* en inglés, que por su propia definición son "cualquier grupo o individuo que puede afectar o ser afectado por la consecución de los objetivos de la empresa" (Argadoña, 1998, pág. 7).

A este respecto son muchos y muy diversos los grupos de interés involucrados en el correcto desarrollo de la actividad empresarial y a los que por tanto les conviene poder anticiparse al riesgo de quiebra. Sin embargo, parece evidente que los grupos de interés más afectados serán los accionistas, tanto presentes como futuros, y los proveedores. Para los accionistas presentes, la relevancia de la quiebra reside en la pérdida del capital invertido además del potencial menoscabo del propio por tener que responder con su patrimonio personal en determinadas situaciones. Para los accionistas futuros, la quiebra puede ser el factor decisivo para invertir o no y determinante en el precio a pagar por la inversión, que también es crítico para prestamistas o financiadores. Para los proveedores, la importancia estriba en que muy probablemente experimentarán una merma en el cobro de sus deudas. Además, dependiendo del tamaño de la empresa proveedora y en función de la cantidad de clientes que esta ostente, puede incluso suponer que la misma quiebre también por falta de solvencia definitiva si la situación se prolonga en el tiempo, generando un efecto cascada.

Otros grupos de interés muy atraídos por conocer el riesgo de quiebra de una empresa podrían ser las comunidades locales, las agencias gubernamentales y las entidades competidoras. Para las comunidades locales y agencias gubernamentales el interés en que una determinada sociedad siga operando es evidente. Su alta inquietud radica en las repercusiones sociales y económicas que conlleva la desaparición de una compañía, entre las que se encuentra el potencial aumento de la tasa de paro o la hipotética necesidad de intervención estatal para asegurar la competencia. En este sentido, las entidades competidoras son unas de las principales interesadas en conocer el riesgo de quiebra de una empresa, porque además de verse beneficiadas por el descenso de competitividad, estas podrán ir un paso más allá, y analizar una oportunidad de negocio.

Cuando se menciona oportunidad de negocio se hace referencia a los supuestos de fusiones y adquisiciones (M&A en adelante por sus siglas del inglés *Mergers & Acquisitions*) por los que una sociedad amplía su tamaño, sus activos o su cartera de clientes. La predicción de quiebras es crucial para identificar compañías con baja solvencia en el contexto de M&A, que pueden ser adquiridas a un precio menor y con mejores condiciones debido a su mermada capacidad de negociación. Esta herramienta ayuda a tomar decisiones informadas y estratégicas en el proceso de adquisición de negocios. Además, con un análisis parejo de la compañía que el modelo predictivo ha clasificado como “quiebra”, se puede completar el motivo del porqué de la quiebra. Una empresa quebrada no siempre indica que sea poco atractiva, ya que diversos factores han podido hacer que sea insolvente, incluyendo una mala gestión. La cartera de clientes, activos tangibles e intangibles o nicho de mercado son factores que pueden hacer muy atractiva a una sociedad a pesar de su insolvencia o riesgo de ella.

En la actualidad, existen muchas empresas dedicadas a aportar información acerca de la solvencia (capacidad de cumplir sus obligaciones financieras a futuro) y liquidez (capacidad de los activos de convertirse en efectivo) de las compañías como “EInforma”, “DatosCif” o “empresia”. No obstante, el objetivo de este análisis es predecir la probabilidad de quiebra en un plazo corto de tiempo, es decir la probabilidad de insolvencia total con un año de antelación. Que una empresa tenga un ratio de liquidez bajo no implica necesariamente una mala gestión, ya que puede deberse a las características del sector concreto. Por ello, es necesario focalizarse en un único sector, para que los ratios financieros puedan compararse entre sí. En este sentido, se va a estudiar la probabilidad de quiebra de empresas de una parte del sector de la restauración en España, concretamente de “restaurantes y puestos de comida”. Entendiendo como tal la prestación de servicios de comida a clientes ya sea mediante servicio o autoservicio y tanto consumida en el local como a domicilio, incluyendo las comidas preparadas para su consumo inmediato adquiridas en carritos o vehículos a motor (Insituto Nacional de Estadística, 2022).

CAPÍTULO 3. ESTADO DEL ARTE

La predicción de quiebras ha sido un tema muy recurrente a lo largo del tiempo. Desde 1930 se han realizado muchos estudios al respecto, principalmente atendiendo al análisis financiero como una cuestión accesoria a la solvencia para predecir la quiebra. Algunos de estos estudios son: Bureau of Business Research, 1930; FitzPatrick, 1932 o Smith & Winakor, 1935 (Bellovary, Giacomino, & Akers, 2007). Hasta mediados de 1960 se analizaban los ratios de manera individual y no es hasta la publicación de "*Financial Ratios As Predictors of Failure*" (Beaver, 1966) que se demostró que los ratios financieros de las empresas que quiebran tienen un patrón estadístico concreto tomados individualmente (Bellovary, Giacomino, & Akers, 2007).

Pocos años más tarde Altman (1968) trató de formular la relación entre la quiebra y los factores empresariales determinantes para ella, con su estudio basado en el análisis discriminante múltiple que se focaliza en cinco ratios financieros para empresas manufactureras. Durante las décadas de los 60 y 70 el método predominante en términos de predicción de quiebra o riesgo de ruina era el análisis discriminante. Mientras que en las décadas de los 80 y 90 con el aumento de la capacidad de procesamiento de los ordenadores, se pasa a estudios basados en inteligencia artificial (Bellovary, Giacomino, & Akers, 2007). Los métodos predominantes para el análisis entonces, eran la regresión logística, en adelante *logit*; implementados por, entre otros, Ohlson (1980), Laitinen & Laitinen (2000) y las redes neuronales; implementadas por O'Leary (1998) o Anandarajan et al. (2001) (Bellovary, Giacomino, & Akers, 2007).

Para poder llevar a cabo un estudio riguroso acerca de la predicción de quiebras es crucial tomar en consideración una amplia variedad de factores, que son; la selección y calidad de las variables y la elección del método predictivo más adecuado, con base estadística o de inteligencia artificial (Collins, 1980; Alaka, y otros, 2018). A continuación, se van a describir las opciones más notorias tanto de selección de variables, como de elección de modelos. Posteriormente se determinará qué variables se van a seleccionar y el criterio que lo justifica y qué métodos predictivos son los más adecuados, y por ende los que se implementarán en el análisis.

3.1. SELECCIÓN DE VARIABLES

La contabilidad es uno de los sistemas de información del desempeño de una empresa más antiguos que existen (Minbiole, Poli, & Haka, 2015). A lo largo de la historia de la predicción de quiebras se han distinguido dos tipos de modelos en función a la información de la que se sirvan para sus predicciones; modelos basados en información contable y no contable.

Por un lado, los modelos basados en información contable facilitan la comparación de empresas entre sí, puesto que se trata de datos estandarizados, fácilmente disponibles, que aportan detalle sobre el desempeño financiero de la empresa en el tiempo. No obstante, los modelos basados exclusivamente en este tipo de datos dejan de lado el contexto del entorno de la empresa, por ejemplo, posibles cambios en las tendencias de la industria o detalles que simplemente no se han visto reflejados todavía en los estados financieros.

Por otro lado, los modelos basados en información no contable incluyen datos macroeconómicos, datos de la industria y otros aspectos relevantes de la organización, como el tamaño empresarial. Estos modelos pueden ser muy útiles para empresas de nueva creación que no han tenido tiempo de consolidar sus partidas contables. Además, la información no contable aporta información relevante acerca del contexto industrial y económico que puede ser pasada por alto si se utiliza información contable únicamente, enriqueciendo así los modelos y potencialmente aumentando su precisión. Sin embargo, además de que conseguir este tipo de información es complicado, la subjetividad inherente a la misma puede ser un aspecto determinante para una baja precisión del modelo.

Es abundante la literatura existente que compara la relevancia de las variables, indicando si es más notable la información contable o la no contable. El estudio de Shumway (2001) demuestra que incorporar variables relativas al entorno empresarial a modelos cuyo núcleo es la información contable, hace que mejoren su precisión significativamente. En la misma línea se pronuncian también Tascón y Castaño (2012). Sin embargo, otros como Jones y Hensher (2008) concluyeron que los factores macroeconómicos no aportaban información significativa al modelo. Con todo, y debido a la dificultad de obtener los datos relativos al contexto económico para cada año, y al amparo de diversos estudios que utilizan información contable únicamente, se va a llevar a cabo un estudio de la predicción de quiebra empresarial del sector de la restauración en España basado en información contable, obtenida de la base de datos SABI.

3.1.1. Selección de ratios contables

La selección de ratios contables ha sido un tema de debate en los estudios académicos llevados a cabo sobre la predicción de quiebras. Como se ha demostrado en diferentes ocasiones, existe contradicción entre los autores acerca de qué ratios son los más indicativos de una quiebra; un ejemplo sería la discusión acerca de la primacía del ratio de liquidez (activo corriente/pasivo corriente) sobre el ratio fondo de maniobra/total activo (Bellovary, Giacomino, & Akers, 2007, pág. 3). Visto que no hay acuerdo acerca de cuáles son los ratios más explicativos de manera genérica, se va a implementar la teoría de los grandes números[2] para escoger las variables *a priori*. Posteriormente se realizará un análisis con el fin de evitar la redundancia de información entre variables y seleccionar las que finalmente serán utilizadas para implementar los modelos predictivos. Sabiendo que en la literatura existen modelos que se basan en el análisis de una sola variable (Beaver, 1966) y otros que se basan en múltiples, el foco de este trabajo de fin de grado se inclina hacia los últimos. En esta línea, se va a hacer referencia a los ratios más utilizados como así se analiza en "*A Review of Bankruptcy Prediction Studies: 1930-Present*" (Bellovary, Giacomino, & Akers, 2007, pág. 42) escogiendo los 20 ratios basados en información contable más utilizados, que son:

[2] Teorema que enuncia que cuantas más veces se repite un suceso más cerca estaremos de encontrar la esperanza de este, y por tanto del comportamiento medio del suceso.

Tabla 1: Ratios financieros utilizados

IDENTIFICACIÓN RATIO	FÓRMULA DEL RATIO	BREVE DESCRIPCIÓN
A	Ingresos netos / Total activo	Medida de eficiencia de los activos para la generación de ingresos.
B	Activo Corriente / Pasivo Corriente	Mide la capacidad de una empresa para hacer frente a sus deudas a corto plazo.
C	Fondo de maniobra / Total Activo	Mide la solvencia de una empresa en el largo plazo, es decir cuánto tiene realmente a corto plazo sobre lo que tiene en total.
D	Patrimonio Neto / Total Activo	Mide la financiación propia de la empresa, que no debería ser superior a 1/3 según los estándares habituales.
E	EBIT [3] / Total Activo	Refleja la rentabilidad de una empresa, es decir qué capacidad tienen sus activos de generar beneficios.
F	Ingresos por Ventas / Total Activo	Mide la eficacia comercial respecto al activo empleados en conseguirlas, y da idea de la rentabilidad comercial que se obtiene del activo.
G	(Activo corriente - Existencias) / Pasivo Corriente	Mide la capacidad de una empresa para hacer frente a los pagos más inmediatos, indicando la capacidad de pagar el pasivo corriente sin recurrir a las existencias.
H	Total Pasivo / Total Activo	Mide el porcentaje de activos de una empresa financiados con deuda externa.
I	Activo Corriente / Total Activo	Mide la proporción de activo que es a corto proveniente de las operaciones.
J	Ingresos Netos / Patrimonio Neto	Refleja la capacidad de generar ingresos en relación con la inversión interna.
K	Tesorería / Total Activo	Nos indica la proporción de recursos dedicada a hacer frente a pagos inmediatos.
L	Flujo de caja de Operaciones / Total Activo	Es una medida de rentabilidad que trata de establecer el ratio de cuán productivo es el capital invertido en el activo.
M	Flujo de caja de Operaciones / Total Pasivo	Mide la capacidad de generar caja respecto al total compromisos de la empresa.
N	Pasivo Corriente / Total Activo	Medida de endeudamiento que refleja qué parte del activo está financiada con deuda externa a corto plazo.
O	Activo Corriente / Ingresos por Ventas	Mide la eficiencia de cobros y tesorería para las ventas que se realizan.
P	EBIT / Intereses	Proporciona la capacidad de la empresa de hacer frente a los gastos financieros.
Q	Existencias / Ingresos por ventas	Mide la eficiencia de la gestión de existencias.
R	Flujo de caja de Operaciones / Ingresos por Ventas	Mide el inverso del margen operativo por unidad de venta.
S	Fondo de maniobra / Ingresos por Ventas	Medida de equilibrio financiero, indica qué parte de los ingresos se destinan al activo circulante.
T	Pasivo Fijo / Total Activo	Medida de endeudamiento que refleja qué parte del activo está financiada con deuda externa a largo plazo.

Fuente: Elaboración propia

[3] Por sus siglas en inglés *Earnings Before Interest & Tax*, en español Beneficio antes de intereses e impuestos.

3.2. SELECCIÓN DE MÉTODO

Como ya se ha mencionado, los métodos predictivos se dividen en dos grandes bloques: los modelos estadísticos y los basados en inteligencia artificial. La diferencia de los segundos radica en que implementan técnicas de aprendizaje integradas a diferencia del mero análisis que llevan a cabo los primeros. Dentro de los modelos estadísticos se encuentran el análisis discriminante simple y la regresión logística. Entre los modelos basados en inteligencia artificial se encuentran, los árboles de decisión, las redes neuronales y las *Support Vector Machines* (máquinas de soporte vectorial, SVM en adelante por sus siglas en inglés). No obstante, es preciso explorar las ventajas e inconvenientes generales de cada enfoque para conocer las limitaciones del estudio.

Siguiendo a Sun, Li, Huang y He (2014) las principales ventajas de los enfoques estadísticos son su fácil implementación, la rapidez y precisión en la obtención de resultados y su utilidad en la identificación de las variables más importantes. Sin embargo, presentan limitaciones en la identificación de patrones complejos y no lineales, así como una elevada sensibilidad a los datos atípicos (*outliers*) y variables poco relevantes, o que aportan poca información.

Por otro lado, los enfoques basados en técnicas de inteligencia artificial pueden identificar patrones complejos y no lineales que resultan difíciles de detectar con otros tipos de modelos. Además, pueden ser más precisos y flexibles en el procesamiento de grandes cantidades de datos. A pesar de ello, estos enfoques pueden ser más complejos de interpretar debido a la falta de explicabilidad de los modelos, que en no pocas ocasiones pueden ser una "caja negra" [4], y son más propensos a sobre ajustarse, es decir, a aprender patrones específicos de los datos de entrenamiento que no se generalizan bien (Sun, Li, Huang, & He, 2014).

[4] Ausencia de explicación en el proceso de la toma de decisión, la ponderación de los factores y los procesos internos son una incógnita.

3.2.1. Métodos basados en estadística

3.2.1.1. Análisis discriminante simple multivariable

En 1968, Altman publicó el estudio "*Financial ratios discriminant analysis and the prediction of corporate bankruptcy*" que serviría de referencia para este tipo de método. En esta investigación se implementó el modelo *z-score,* que con el tiempo cobraría una relevancia vital como la tiene ahora y logra probar que con la información de tan solo cinco ratios [5] se puede lograr una tasa de acierto del 95% con un año de antelación.

El análisis discriminante simple multivariable [6] es una técnica estadística que consiste en maximizar la varianza relativa de la intersección de los dos grupos predefinidos ("quiebra" o "no quiebra") en función de uno de ellos, obteniendo la combinación lineal de las variables independientes. La combinación obtenida, o *z-score* de cada empresa, se compara después con un valor de corte que es el factor determinante para clasificar a la empresa en uno de los grupos (Altman, 1968). Se explica de la siguiente forma:

$$Z = w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

Donde:

Z = Puntuación discriminante.

w_n = Pesos discriminantes

x_n = Ratios financieros

n = Número de ratios

Los requisitos para que este modelo funcione adecuadamente incluyen que las variables sigan una distribución Normal, que las matrices de covarianza sean iguales para cada grupo y que los grupos no se superpongan. No obstante, otro factor que hace difícil la aplicación práctica del modelo es la multicolinealidad entre las variables a *priori* independientes (Back, Laitinen, Sere, & van Wezel, 1996).

[5] Que son: Fondo de maniobra/Activo Total; Ganancias Retenidas/Activo Total; EBIT/Activo total; Valor de mercado del Patrimonio Neto/Valor en libros de Pasivo Total; Ingresos por ventas/Activo Total.

[6] Realiza la clasificación en dos grupos, es decir binaria, a diferencia del compuesto, que divide entre más de dos clases.

3.2.1.2. Regresión logística – Logit

La regresión logística es un modelo de probabilidad condicional que, bajo la asunción de una distribución logística, estima la probabilidad de quiebra de una empresa. Esto lo hace utilizando el criterio de la máxima verosimilitud logarítmica, que trata de maximizar la probabilidad de que los datos observados se ajusten a la distribución del modelo. Trata de predecir el resultado de una variable categórica ("quiebra", "no quiebra") en función de las variables independientes (ratios financieros). De esta forma no es necesario establecer un valor de corte determinado para establecer la frontera de división entre empresa que quiebra y empresa que no, como sí lo es en el análisis discriminante. A diferencia de este, la regresión logística asigna a cada empresa, que se representa como una fila en la base de datos, una probabilidad de quiebra en función del nivel de confianza predispuesto.

Aplicando la regresión logística, fue Shumway (2001) con su modelo Hazard, quien amplió el alcance de los datos utilizados para la predicción de quiebras. Esto lo hizo incluyendo el transcurso del tiempo como una variable dependiente, que representa la covarianza de los datos en el tiempo. Con esta función se clasifican las empresas en quebradas o no quebradas para cada año, siendo 0 no quiebra y 1 quiebra, asociando una probabilidad diferente a cada año en función de la clasificación.

La principal ventaja que aporta el modelo de regresión logística es la independencia de presunciones *a priori* de la distribución de probabilidad de las variables, ya que en principio los ratios financieros no se adaptan a la distribución normal (Rodríguez, Piñeiro, & de Llano). Además, con el modelo *logit* se obtiene información muy relevante acerca de la importancia de las variables. Al ser un método explicativo es muy útil para extraer conclusiones.

3.2.2. Métodos basados en inteligencia artificial

3.2.2.1. Árboles de decisión

Existen diversas técnicas basadas en la inteligencia artificial, entre las que se pueden encontrar los árboles de decisión. Los árboles de decisión son una técnica predictiva de aprendizaje supervisado que, de manera jerárquica hacen división entre las variables independientes (ratios financieros) y establecen una partición (por ejemplo, que el ratio de solvencia sea mayor que 0,5) para llegar a clasificar la empresa como quebrada o no. La partición dentro de cada regla se establece en función de cuál de todos los posibles valores de la división hace disminuir la impureza (error de clasificación) dentro de las clases a las que da lugar. De esta

manera se van ordenado las variables independientes en función de la información que aportan a cada partición para clasificar en quebrada o no quebrada a una empresa.

Es un método bastante eficaz para predecir quiebras puesto que es capaz de modelar relaciones no lineales y su resultado es fácil de interpretar (Quinlan, 1986; Kwon & Cho, 2016). No obstante, este método tiende al sobreajuste en casos como el que se trata aquí, que cuentan con una elevada cantidad de variables independientes. Además, es muy inestable, ya que pequeños cambios en el conjunto de datos de entrada darán lugar a un árbol diferente, que no se podrá garantizar como el óptimo en ningún caso, ni aun usando las técnicas de *pruning* o *random forest* para mejorarlo. Esto se debe en parte a que el orden jerárquico de las variables y posteriores divisiones son en gran medida heurísticos, dependiendo del orden que se siga al analizarlas (Jeng, Jeng, & Liang, 1997).

3.2.2.2. Redes neuronales

Este método tuvo su auge en las décadas de 1980 y 1990 con el desarrollo de las tecnologías y el aumento de la memoria en los ordenadores. De todas las herramientas de inteligencia artificial, las redes neuronales y las redes neuronales artificiales son las que más se han utilizado para resolver el problema de la predicción de quiebras (Aziz & Dar, 2006; Tseng & Hu, 2010). De esta forma se trata de un método de clasificación binaria que mediante la creación de conexiones entre sus "neuronas" dentro de las distintas capas, encuentra patrones en los datos proporcionados. La red neuronal tiene tantas neuronas en su primera capa como variables, pasando esta información a varias neuronas de la siguiente capa, aunque cada neurona solo emite una salida que servirá de entrada para varias neuronas de la siguiente capa. Así es como las neuronas aprenden de los datos analizando sus patrones a diferentes niveles para después otorgar la calificación de "quiebra" o "no quiebra" a una empresa (Agarwal, 1993).

La principal ventaja que arrojan las redes neuronales es la capacidad de tratar con una elevada cantidad de datos y variables puras, es decir sin necesidad de pretratar o normalizar (Agarwal, 1993). Sin embargo, uno de los principales inconvenientes que este método brinda es la intensidad computacional y la ausencia de explicación teórica por la dificultad de entender los modelos matemáticos subyacentes del algoritmo, actuando, así como un modelo "caja negra", siendo imposible predecir su resultado (Alaka, y otros, 2018)

3.2.2.3. Support Vector Machines

Otro método que ha sido utilizado en múltiples ocasiones para predecir la quiebra empresarial a los largo de los años ha sido el SVM (Alaka, y otros, 2018). Es un método de clasificación binaria que representa las variables independientes en un hiperplano en una dimensión menor, buscando la mayor distancia entre este y la variable objetivo, en este caso quebrar o no. El objetivo de este método es encontrar el hiperplano óptimo de entre todos los posibles, es decir, el que menos errores cometa en la clasificación.

Este método tiene una alta capacidad para manejar grandes conjuntos de datos y asimilar relaciones no lineales (Alaka, y otros, 2018). Sin embargo, cuantas más variables se utilicen y más redundancia exista entre ellas, menos capaz será de hacer una buena predicción (Raj, 2022).

3.2.3. Elección de métodos a implementar

Con todo, muchos son los estudios que comparan las técnicas de predicción de quiebras en base a su explicabilidad o su precisión. Dentro de los modelos basados en técnicas estadísticas, los estudios comparativos parecen concluir que el análisis discriminante y la regresión logística son dentro de los modelos estadísticos los más robustos para predecir la quiebra empresarial (Alaka, y otros, 2018; Bellovary, Giacomino, & Akers, 2007). De entre los dos, en este trabajo de fin de grado se ha optado por la regresión logística (*logit*) por ser más explicable y exigir menos requisitos, además de su facilidad de implementación computacional en los paquetes preestablecidos en Python (librerías).

Por otro lado, dentro de los modelos de inteligencia artificial destacan las redes neuronales por encima del resto de métodos por su precisión a la hora de medir la quiebra empresarial, que incluso lo anteponen a otros métodos como los árboles de decisión (Alaka, y otros, 2018; Bellovary, Giacomino, & Akers, 2007; Iturriaga & Sanz, 2015). Además, según la investigación de Zhang, Li y Guo (2019), las redes neuronales tienen una mayor capacidad para modelar relaciones complejas y capturar patrones no lineales en comparación con los métodos de SVM y árboles de decisión. Esto constituye el principal motivo de la elección de las redes neuronales entre todas las técnicas de inteligencia artificial para ser implementada en este trabajo de fin de grado.

CAPÍTULO 4. APLICACIÓN EMPIRICA

4.1. DATOS

Para llevar a cabo la investigación propuesta, es necesario descargar la información financiera de las empresas del sector elegido. Para este propósito se cuenta con una suscripción a la base de datos SABI. Esta base de datos contiene información del histórico de datos contables reportados por las entidades españolas. Con el objetivo de predecir la quiebra empresarial con un año de antelación, se ha optado por descargar el último año disponible de cada empresa y el año anterior a este en términos absolutos. No se ha optado por años determinados, que SABI denomina "años relativos" (por ejemplo 2015-2020) con la intención de aumentar el número de registros, sin acotar temporalmente el *dataset*. Teniendo esto en cuenta se obtendrá un conjunto de datos que de la opción a predecir si una empresa va a quebrar o no con un año de antelación.

La clasificación por sectores que sigue SABI es la CNAE-2009 (de sus siglas Clasificación Nacional de Actividades Económicas), impuesta por la Unión Europea para fines estadísticos. Esta clasificación va desgranando los sectores empresariales. Dentro del sector "5. Hostelería" se pueden encontrar diferentes subapartados, el que interesa para la realización de este trabajo es "56. Servicios de comidas y bebidas", que incluye las actividades de prestación de servicios de comidas y bebidas listas para su consumo inmediato independientemente de las instalaciones con las que cuente. En este apartado se incluyen diferentes subpartidas que son; "561. Restaurantes y puestos de comidas", "562. Provisión de comidas preparadas para eventos y otros servicios de comidas" y "563. Establecimientos de bebidas".

Es importante entender qué incluye cada epígrafe para poder seleccionar un sector comparable. El epígrafe escogido, 561, está definido como: la prestación de servicios de comida a clientes ya sea mediante servicio o autoservicio y tanto consumida en el local como a domicilio, incluyendo las comidas preparadas para su consumo inmediato adquiridas en carritos o vehículos a motor. Sin embargo, excluye el comercio mediante máquinas expendedoras y la explotación de concesiones del servicio de restauración. El epígrafe 562 hace referencia a las actividades de provisión de comidas preparadas para eventos o de una duración determinada y el 563, comprende la preparación y el servicio de bebidas para el consumo inmediato en el mismo local, donde la actividad predominante sea servir bebidas. (Insituto Nacional de Estadística, 2022)

El motivo de elección de la categoría "restaurantes y puestos de comida" y excluir tanto los servicios de preparación de eventos como los establecimientos de bebidas, es que la preparación de alimentos y la de bebidas tiene distinto margen de beneficio. Esto se debe principalmente a los gastos involucrados, aunque también a las diferentes formas de gestionar la

mercancía, pues los alimentos tienen un ciclo de vida más corto que las bebidas. Con respecto a los servicios para preparación de eventos, no se pueden incluir por la estacionalidad de estos, que genera diferentes tendencias a la hora de administrar tanto mercancía, como patrimonio. Por tanto, a pesar de ser industrias relacionadas, no son comparables para el propósito de este trabajo.

SABI contiene 62.127 registros de empresas asignadas a la categoría “restaurantes y puestos de comida”, que son los que se van a emplear para el análisis. De cada empresa se han descargado los campos descritos en la tabla 2 para poder elaborar los ratios financieros que servirán de variables independientes de entrada para el modelo.

Tabla 2: Descripción de las variables contables descargadas

CAMPO	VALORES QUE TOMA	DESCRIPCIÓN	ESTADO FINANCIERO AL QUE PERTENECE
CIF	Texto	Código de Identificación Fiscal único en España	-
Localización	Texto	Municipio donde está localizada la empresa	-
Último año disponible	Fecha	Fecha de último año disponible	-
Activo circulante	Cantidad en miles de euros	Hace referencia al activo realizable en el corto plazo	Balance de situación
Existencias	Cantidad en miles de euros	Corresponde al valor de los activos para el proceso de producción/venta	Balance de situación
Deudores	Cantidad en miles de euros	Corresponde a las deudas de la que la empresa es acreedora que tienen un plazo de vencimiento menor a un año	Balance de situación
Tesorería	Cantidad en miles de euros	Efectivo o dinero en cuentas corrientes neto	Balance de situación
Total activo	Cantidad en miles de euros	Corresponde a la fórmula: activo circulante + activo fijo/ largo plazo	Balance de situación
Fondos propios	Cantidad en miles de euros	Corresponde a las aportaciones de los socios de la empresa y sus beneficios, no a financiación externa	Balance de situación
Fondo de maniobra	Cantidad en miles de euros	Representa la capacidad de afrontar las deudas a corto plazo de una empresa, si es positivo el valor se trata de una empresa solvente y si es negativo una que no lo es	Balance de situación
Total pasivo y capital propio	Cantidad en miles de euros	Valor total financiado, debe ser igual al activo	Balance de situación
Pasivo líquido	Cantidad en miles de euros	Hace referencia a las deudas que la empresa debe pagar en un plazo menor a un año	Balance de situación
Pasivo fijo	Cantidad en miles de euros	Hace referencia a las deudas que la empresa debe pagar en un plazo mayor a un año	Balance de situación
Ingresos de explotación	Cantidad en miles de euros	Simbolizan las entradas monetarias obtenidas por la empresa consecuencia de su actividad ordinaria	Cuenta de Pérdidas y Ganancias
Resultado de Explotación	Cantidad en miles de euros	Corresponde a la resta de ingresos de explotación menos gastos provenientes de la actividad ordinaria	Cuenta de Pérdidas y Ganancias
EBIT	Cantidad en miles de euros	Beneficio antes de intereses e impuestos	Cuenta de Pérdidas y Ganancias
Gastos financieros	Cantidad en miles de euros	Representa los gastos derivados de la financiación externa de la empresa (ej. Intereses, comisiones…)	Cuenta de Pérdidas y Ganancias
Flujo de caja	Cantidad en miles de euros	Hace referencia al conjunto de flujos de caja de una empresa, de entrada y salida	Estado de flujos de caja

Fuente: Elaboración propia

Hubiera sido interesante descargar todos los registros relativos al balance de situación y la cuenta de pérdidas y ganancias, sobre todo lo referente al resultado del ejercicio. Sin embargo, debido a una limitación establecida por la web no se ha podido descargar más información. Otra limitación a la que ha habido que enfrentarse es que algunas partidas no existían, por ejemplo, inventario. Por tanto, para llevar a cabo el ratio financiero que requería del inventario, se ha suplido este por la partida de existencias, que es la que contiene la información más similar disponible. Si fuera posible contar con todo el detalle real de las partidas contables se podría establecer una comparativa más fidedigna con los estudios previos a la hora de otorgarle un poder de predicción a cada ratio financiero.

4.1.1. Pretratamiento

En primer lugar, se ha creado la variable objetivo, es decir la variable quiebra, donde se ha otorgado un valor "0" a las empresas que no quiebran y se ha asignado un valor "1" a las que si lo hacen. La base de datos origen no contaba con esta información, sino que tenía al final, dentro del campo "empresa" el contenido "extinguida" o "en liquidación" según en qué fase del concurso de acreedores se encontrara, si es que estaba incursa en uno (Ej: "GRUPO ZENA DE RESTAURANTES SA (EXTINGUIDA)"). Por ello con las reglas de Excel y sus funciones integradas[7] se ha rellenado la columna "quiebra" de manera automática.

En segundo lugar, también en Excel, se han obtenido los ratios financieros que servirán de variables independientes para los modelos que van a ser implementados. Se ha establecido una leyenda para nombrar los ratios, ya que, que el nombre de identificación de una variable contenga símbolos no alfanuméricos no es recomendable y no todos los ratios cuentan con un nombre común. La leyenda nombra los ratios de la letra "A" a la letra "T" del abecedario para la información relativa al último año disponible, es decir el año *n*. De la misma forma, se ha establecido de la letra "A" a la letra "T", seguido de un "1" (Ej: "A1"), para denominar la información del año anterior al último disponible, es decir la información del año *n-1*. Las fórmulas que componen cada ratio, que se explicitaron en el epígrafe 3.1.1. se configuran de la manera que aparece en la tabla siguiente.

[7] =SI (SI. ERROR (HALLAR("EXTINGUIDA"; B54;1);0)+SI.ERROR(HALLAR("liquidacion";B54;1);0)>0;1;0)

Tabla 3: Ratios financieros asociados con cada variable

VARIABLE	DESCRIPCIÓN
A	Ingresos netos (n) / Total activo (n)
A1	Ingresos netos (n-1) / Total activo (n-1)
B	Activo Corriente (n) / Pasivo Corriente (n)
B1	Activo Corriente (n-1) / Pasivo Corriente (n-1)
C	Fondo de maniobra (n) / Total Activo (n)
C1	Fondo de maniobra (n-1) / Total Activo (n-1)
D	Patrimonio Neto (n) / Total Activo (n)
D1	Patrimonio Neto (n-1) / Total Activo (n-1)
E	EBIT (n) / Total Activo (n)
E1	EBIT (n-1) / Total Activo (n-1)
F	Ingresos por Ventas (n) / Total Activo (n)
F1	Ingresos por Ventas (n-1) / Total Activo (n-1)
G	(Activo corriente - Existencias) (n) / Pasivo Corriente (n)
G1	(Activo corriente - Existencias) (n-1) / Pasivo Corriente (n-1)
H	Total Pasivo (n) / Total Activo (n)
H1	Total Pasivo (n-1) / Total Activo (n-1)
I	Activo Corriente (n) / Total Activo (n)
I1	Activo Corriente (n-1) / Total Activo (n-1)
J	Ingresos Netos (n) / Patrimonio Neto (n)
J1	Ingresos Netos (n-1) / Patrimonio Neto (n-1)
K	Tesorería (n) / Total Activo (n)
K1	Tesorería (n-1) / Total Activo (n-1)
L	Cash Flow de Operaciones (n) / Total Activo (n)
L1	Cash Flow de Operaciones (n-1) / Total Activo (n-1)
M	Cash Flow de Operaciones (n) / Total Pasivo (n)
M1	Cash Flow de Operaciones (n-1) / Total Pasivo (n-1)
N	Pasivo Corriente (n) / Total Activo (n)
N1	Pasivo Corriente (n-1) / Total Activo (n-1)
O	Activo Corriente (n) / Ingresos por Ventas (n)
O1	Activo Corriente (n-1) / Ingresos por Ventas (n-1)
P	EBIT (n) / Gastos financieros (n)
P1	EBIT (n-1) / Gastos financieros (n-1)
Q	Existencias (n) / Ingresos por ventas (n)
Q1	Existencias (n-1) / Ingresos por ventas (n-1)
R	Cash Flow de Operaciones (n) / Ingresos por Ventas (n)
R1	Cash Flow de Operaciones (n-1) / Ingresos por Ventas (n-1)
S	Fondo de maniobra (n) / Ingresos por Ventas (n)
S1	Fondo de maniobra (n-1) / Ingresos por Ventas (n-1)
T	Pasivo Fijo (n) / Total Activo (n)
T1	Pasivo Fijo (n-1) / Total Activo (n-1)

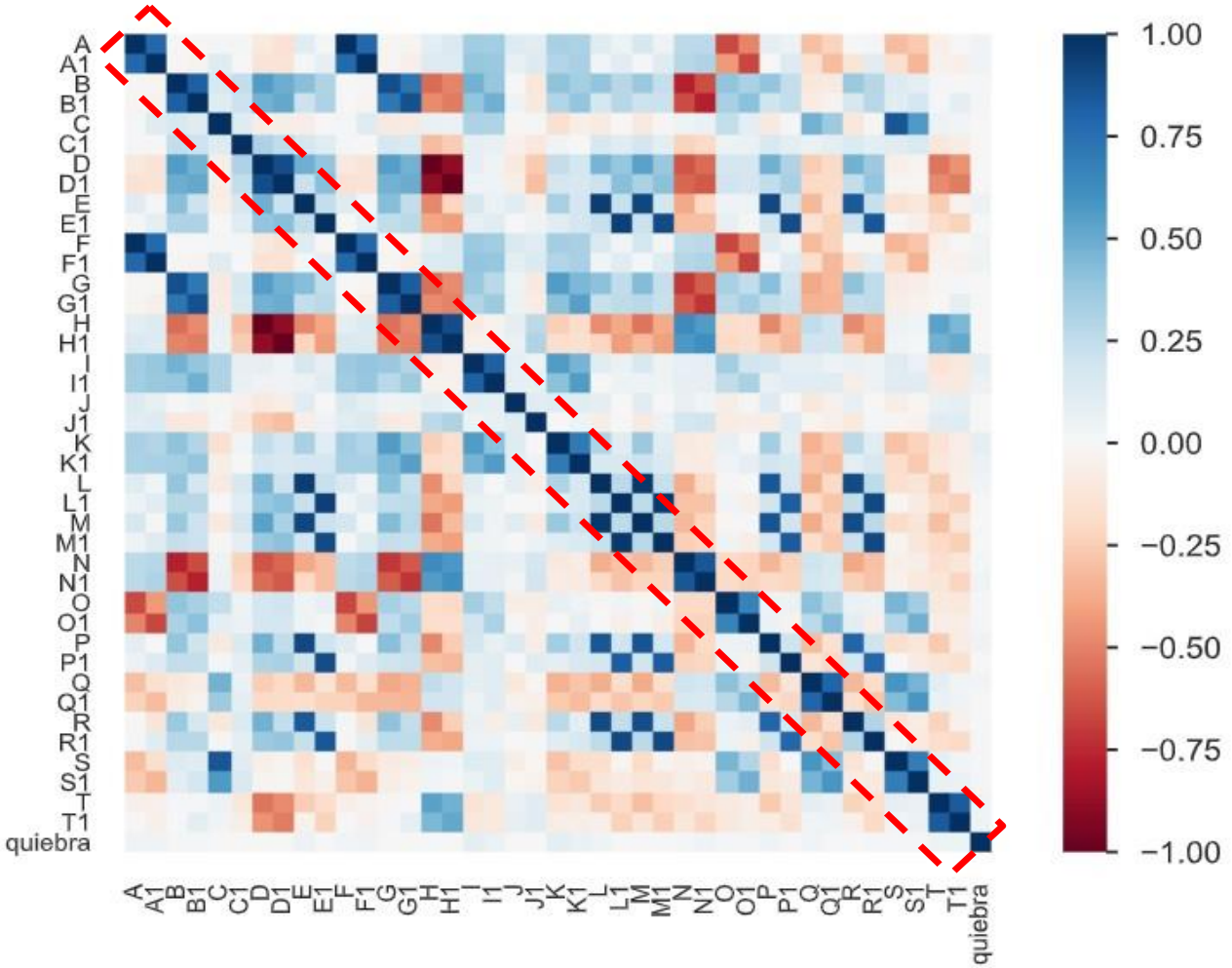
Fuente: Elaboración propia

A continuación, se va a realizar un análisis exploratorio de los datos disponibles con el fin de optimizar la muestra que se va a utilizar para la implementación de los modelos predictivos. Para ello se van a analizar las correlaciones existentes entre las variables y explorar diferentes métodos para atenuarlas. También se va a analizar la densidad de datos de la muestra, examinando los valores no disponibles y qué hacer al respecto.

4.1.1.1. Análisis exploratorio de la muestra y enriquecimiento del dataset

Se ha realizado un análisis previo de la muestra para posteriormente proceder a la implementación del modelo, teniendo en cuenta los factores que de este análisis se extraen. En este sentido se ha obtenido la matriz de correlación que se puede ver en la figura 1, que muestra que en efecto existe una alta correlación en línea generales entre las variables, es decir, entre un ratio y su correspondiente del año anterior.

Figura 1: Matriz de correlación de las variables originales



Fuente: Elaboración propia

Al observar tanta multicolinealidad, y sabiendo que los modelos con redundancia de información tienden a un peor desempeño en las predicciones, se ha optado por crear una nueva variable para tratar de evitarlo. La variable creada, ha sido denominada con las mismas letras de la "A" a la "T", seguida de un R de ratio (Ej: "AR"). El ratio al que hace referencia esta letra es el resultado de dividir el ratio de *n* entre el de *n-1*. Se ha optado por el cociente de variables

"independientes" por varios motivos, inspirado en la creación de variables del estudio de Shumway (2001). El primero, es que no se puede eliminar ninguna de las dos variables como tal, ni *n*, ni *n-1*, pues entonces se perdería la información que da sentido al análisis, predecir la quiebra con un año de antelación. El segundo motivo, es que otros tipos de operaciones matemáticas, como la diferencia tienen en cuenta las magnitudes, es decir terminan por implementar un sesgo en cuanto a cantidades. De esta forma, el cociente, es la manera de mantener toda la información de *n* con respecto a *n-1* minimizando la colinealidad entre las variables y el sesgo. Por tanto, se tienen en cuenta todas las variables independientes y al implementar los modelos, se explorará cuál es la combinación de variables que da mejor resultado en los modelos; si las variables originales de año y año anterior o la de ratios y datos del año anterior. Las variables que servirán de entrada al modelo quedan resumidas en la tabla siguiente:

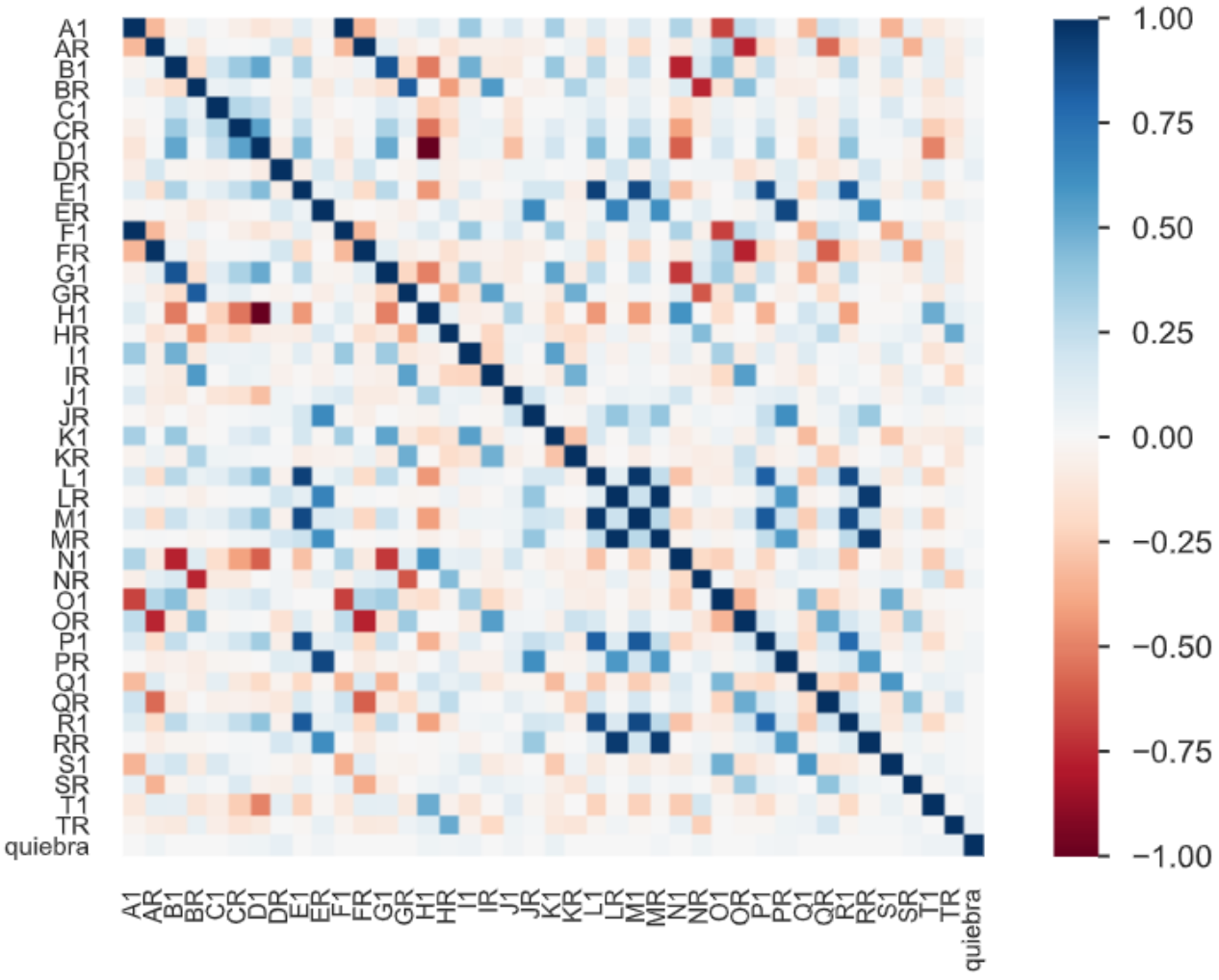
Tabla 4:Variables independientes input de los modelos

VARIABLE	DESCRIPCIÓN	VARIABLE	DESCRIPCIÓN
A	Ingresos netos (n) / Total activo (n)	K	Tesorería (n) / Total Activo (n)
A1	Ingresos netos (n-1) / Total activo (n-1)	K1	Tesorería (n-1) / Total Activo (n-1)
AR	A/A1	KR	K/K1
B	Activo Corriente (n) / Pasivo Corriente (n)	L	Cash Flow de Operaciones (n) / Total Activo (n)
B1	Activo Corriente (n-1) / Pasivo Corriente (n-1)	L1	Cash Flow de Operaciones (n-1) / Total Activo (n-1)
BR	B/B1	LR	L/L1
C	Fondo de maniobra (n) / Total Activo (n)	M	Cash Flow de Operaciones (n) / Total Pasivo (n)
C1	Fondo de maniobra (n-1) / Total Activo (n-1)	M1	Cash Flow de Operaciones (n-1) / Total Pasivo (n-1)
CR	C/C1	MR	M/M1
D	Patrimonio Neto (n) / Total Activo (n)	N	Pasivo Corriente (n) / Total Activo (n)
D1	Patrimonio Neto (n-1) / Total Activo (n-1)	N1	Pasivo Corriente (n-1) / Total Activo (n-1)
DR	D/D1	NR	N/N1
E	EBIT (n) / Total Activo (n)	O	Activo Corriente (n) / Ingresos por Ventas (n)
E1	EBIT (n-1) / Total Activo (n-1)	O1	Activo Corriente (n-1) / Ingresos por Ventas (n-1)
ER	E/E1	OR	O/O1
F	Ingresos por Ventas (n) / Total Activo (n)	P	EBIT (n) / Gastos financieros (n)
F1	Ingresos por Ventas (n-1) / Total Activo (n-1)	P1	EBIT (n-1) / Gastos financieros (n-1)
FR	F/F1	PR	P/P1
G	(Activo corriente - Existencias) (n) / Pasivo Corriente (n)	Q	Existencias (n) / Ingresos por ventas (n)
G1	(Activo corriente - Existencias) (n-1) / Pasivo Corriente (n-1)	Q1	Existencias (n-1) / Ingresos por ventas (n-1)
GR	G/G1	QR	Q/Q1
H	Total Pasivo (n) / Total Activo (n)	R	Cash Flow de Operaciones (n) / Ingresos por Ventas (n)
H1	Total Pasivo (n-1) / Total Activo (n-1)	R1	Cash Flow de Operaciones (n-1) / Ingresos por Ventas (n-1)
HR	H/H1	RR	R/R1
I	Activo Corriente (n) / Total Activo (n)	S	Fondo de maniobra (n) / Ingresos por Ventas (n)
I1	Activo Corriente (n-1) / Total Activo (n-1)	S1	Fondo de maniobra (n-1) / Ingresos por Ventas (n-1)
IR	I/I1	SR	S/S1
J	Ingresos Netos (n) / Patrimonio Neto (n)	T	Pasivo Fijo (n) / Total Activo (n)
J1	Ingresos Netos (n-1) / Patrimonio Neto (n-1)	T1	Pasivo Fijo (n-1) / Total Activo (n-1)
JR	J/J1	TR	T/T1

Fuente: Elaboración propia

Se puede observar en la figura 2 que en efecto las nuevas variables, que hacen referencia al ratio combinado de *n* y *n-1*, están menos relacionadas con respecto a *n-1*, que lo que estaban las variables entre sí, como se ve en la figura 1, es decir, recogen menos colinealidad entre sí. Además, como se pone de manifiesto en la misma figura, la nueva matriz de correlación muestra una multicolinealidad entre variables baja, por lo que será el modelo el que tenga que establecer las conexiones y los patrones que siguen los datos. *A priori*, parece sorprendente, pues al final se trata de información dependiente entre sí, ya que la mayoría de ratios se han elaborado a partir de partidas de la cuenta de pérdidas y ganancias y del balance de situación. En este sentido, si se observa la correlación de las variables independientes con la variable objetivo, "quiebra", se observa que la correlación es baja, así que será complejo construir un modelo de alta capacidad predictiva. Como se hizo referencia en el estado del arte, las redes neuronales son capaces de identificar relaciones más complejas entre los datos, por lo que parece previa implementación que será más preciso este modelo.

Figura 2: Matriz de correlación de las variables n-1 y ratio



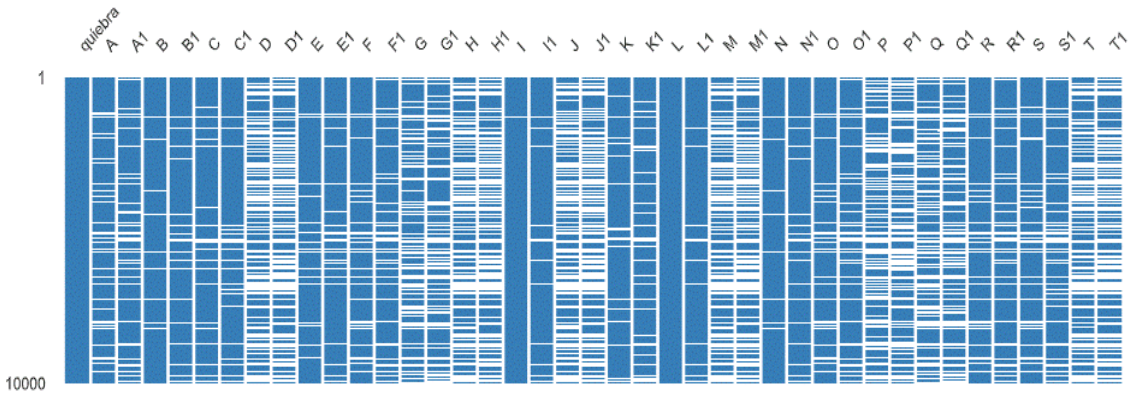
Fuente: Elaboración propia

De los 62.127 registros con los que cuenta la muestra inicial, tan solo 8.872 corresponden a empresas quebradas. Es decir, la quiebra solo afecta a un 14,2% de las empresas con las que se cuenta para la investigación. Es importante saber que el porcentaje de una de las clases, quiebra, es mucho más bajo que el de la otra, no quiebra, puesto que denota desequilibrio muestral y debe tenerse en cuenta a la hora de implementar los modelos. El principal peligro de las muestras desequilibradas es que pueden dar lugar a niveles altos de precisión (*accuracy* comúnmente denominado en inglés) que no han aprendido realmente de la información proporcionada por los datos, sino que se basan en la mera probabilidad en términos generales. Es decir, si el modelo predice que una empresa no va a quebrar, acertará en el 85,8% de los casos. Por tanto, se estaría ante un nivel de precisión muy alto, pero que no aporta valor, puesto que no ha aprendido de los datos proporcionados. Puede ser más interesante un nivel de precisión más bajo pero que realmente aporte información acerca de la propensión de una empresa a quebrar o no.

4.1.1.2. Datos omitidos o no disponibles

Inicialmente se contaba con 62.127 registros de empresas, sin embargo, tras observar la muestra detenidamente se ha visto que hay gran cantidad de valores no disponibles (NA, por sus siglas en inglés de *Not Available*) para diferentes variables. No obstante, la mayoría se concentran en las variables "D", "H", "J", "M", "P", "Q", "T" y sus correspondientes de *n-1*, como se puede observar en la figura 3. La figura 3 muestra la densidad de información contenida en cada variable, siendo los huecos en blanco los correspondientes a valores perdidos, en la figura 4 se puede ver la densidad de cada variable, así como la cantidad de registros no perdidos en los 10.000 primeros registros, por limitaciones de computación.

Figura 3: Mapa de calor de valores perdidos

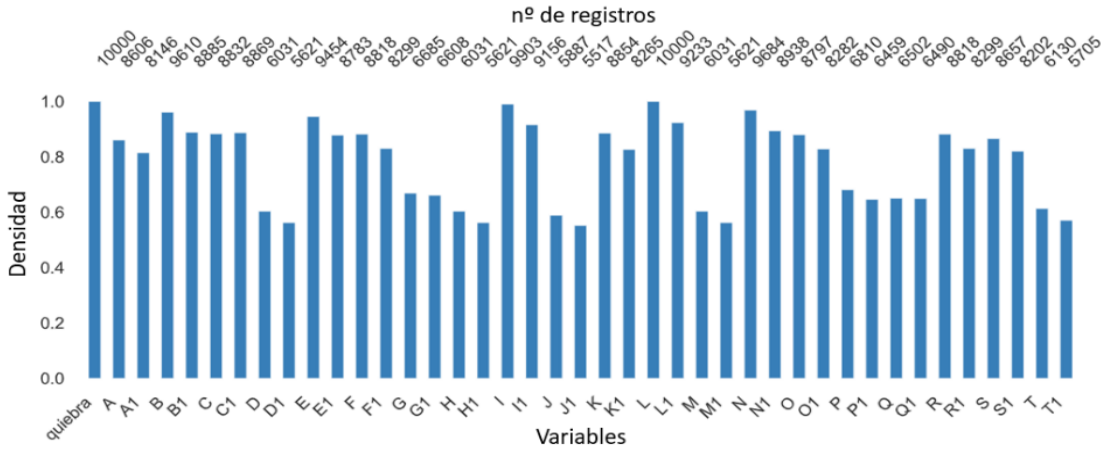


Fuente: Elaboración propia

Es bien sabido que los modelos de *machine learning* no pueden procesar registros en los que no existe información para una variable concreta, por tanto, se debe solucionar este problema antes de proceder a la implementación de los modelos. Existen diferentes maneras de tratar a los valores perdidos o no disponibles, y una de ellas podría ser imputar valores de la media de los no perdidos. No obstante, en este caso no es una opción recomendable, dado que en el set de datos existen empresas de muy diferentes tamaños, cuya contabilidad está muy polarizada, por tanto, sería imputar valores probablemente más altos de los que se obtendrían realmente, alejando el set de datos de la realidad, puesto que no existe un perfil de empresa medio.

En este caso, se procede a la eliminación de los valores perdidos. Como se puede observar en la figura 4 la muestra cuenta con variables que contienen una densidad baja de información, en torno al 0,5. A pesar de que la figura solo analice los primeros 10.000 resultados, al extrapolarlo al conjunto total se ha observado que la muestra cuenta con 44.746 registros que contienen valores perdidos para alguna de las variables, por tanto, tras eliminarlos se ha disminuido bastante la muestra, en torno a un 60%, obteniendo 17.381 resultados. Para optimizar la pérdida de registros se podría determinar para qué variables existen más valores perdidos, como se observa en las figuras 3 y 4 y eliminarlas, haciendo que el tamaño de la muestra disminuya menos, pero perdiendo información financiera acerca de las empresas. Como se ha analizado en este mismo epígrafe, para optimizar en este caso habría que eliminar las variables "D", "H", "J", "M", "P", "Q", "T" y sus correspondientes de *n-1*, pues son las que más valores perdidos presentan, y posteriormente eliminar el resto de valores perdidos.

Figura 4: Densidad de información de las variables



Fuente: Elaboración propia

De los 17.381 registros que van a ser utilizados como muestra para el modelo, 1.714 corresponden a empresas en quiebra y 15.667 a empresas sanas. Esto denota que un 9,8% de empresas quiebran de los datos disponibles, por tanto, se obtiene un amuestra más desequilibrada

que la que se tenía previamente (14,2% de empresas quebradas). Siendo esto así hay que, por los motivos ya explicados, prestar especial atención al balanceo de la muestra para la implementación de ambos modelos, tanto la regresión logística, como la red neuronal. Hay que destacar también que se podría concluir a raíz de esto que no disponer de determinados datos contables no es señal de mayor probabilidad de quiebra.

4.1.2. Equilibrio de la muestra

Equilibrar o balancear la muestra significa establecer un número similar de registros asignados al "0" y al "1" de la variable quiebra, tanto en el conjunto de datos de entrenamiento como en el conjunto de datos de test. Los datos de entrenamiento son los utilizados para que el modelo aprenda y los de test para comprobar los resultados de la implementación del modelo concreto, lógicamente cuantos más datos sean empleados, mejor aprenderá el modelo.

El equilibrio de la muestra es vital para comprender los resultados de un modelo predictivo, pues puede tener un nivel de precisión muy alto, aún sin haber establecido relaciones entre las variables independientes. Para el análisis que se lleva a cabo en este trabajo, esto es de especial importancia, pues si el modelo predijera en todas las ocasiones que una empresa no quebraría, acertará en el 90%. Sin embargo, lo que interesa para empresas que están buscando adquirir otras es la propensión a la quiebra, no que necesariamente vayan a quebrar, por tanto, no es tan relevante el nivel de acierto sino los motivos que hacen que una empresa esté mal clasificada como quebrada. Es decir, resultarán interesantes para un posterior análisis exhaustivo las empresas que el modelo califique como quebradas que no lo sean, pues puede que lo hagan en un futuro o que su liquidez sea muy baja y por tanto sean más proclives a aceptar condiciones menos favorables de fusión con tal de no caer en esta situación. Estas empresas mal clasificadas también pueden ser el reflejo de situaciones concursales previas, ya explicadas en el epígrafe 2.2., como la fase de convenio o la fase común, que no aparecen reflejadas en la base de datos, si bien son relevantes para el caso, pues indican que una empresa se encuentra en situación patrimonial de concurso.

Siendo el balanceo una necesidad de las muestras tan heterogéneas como la que se trata aquí, existen diferentes maneras de llevarlo a cabo. En primer lugar, podría realizarse sirviéndose de técnicas de creación de información artificial, implementando la función "smote()", que trata de crear datos sintéticos a partir de los existentes de la clase con carencia de registros, la quiebra en este caso. En segundo lugar, se puede advertir al modelo que la muestra esta desbalanceada, para que sea el modelo quien optimice el modelo, pasándoselo como advertencia en los hiperparámetros al llamar a la función integrada, que es lo que se realizará para la regresión

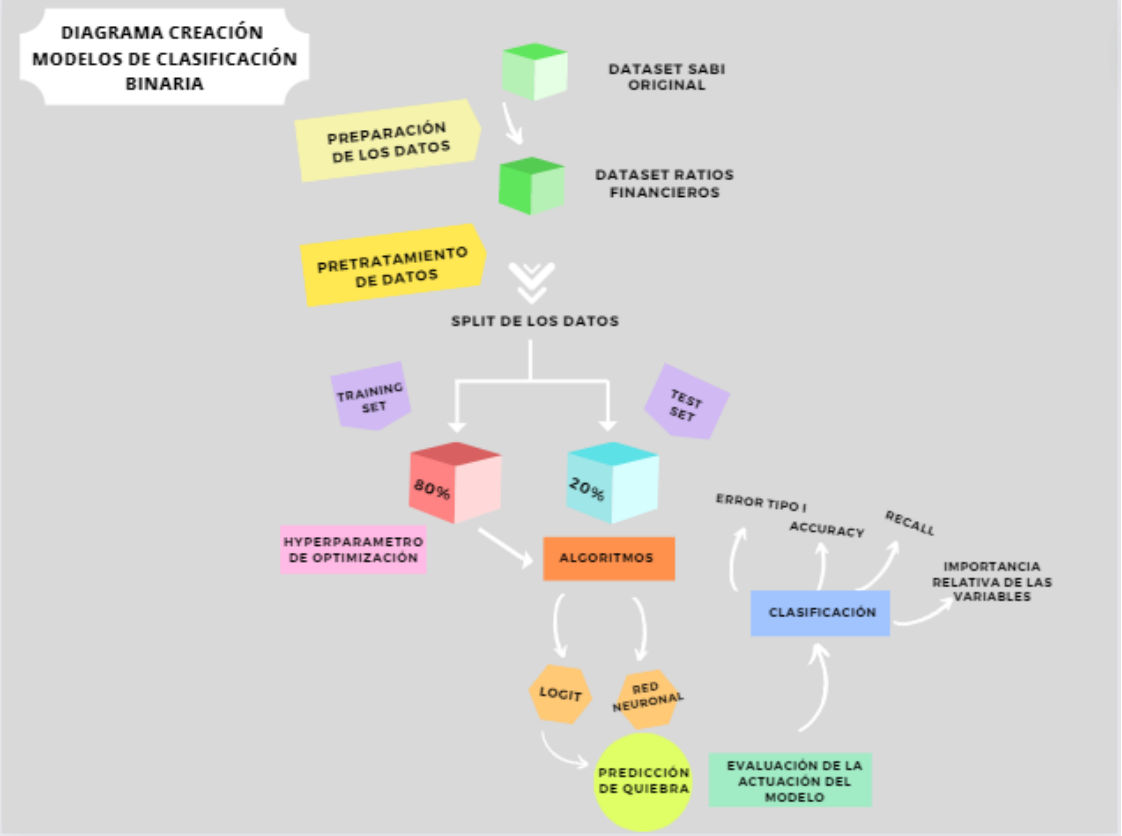
logística, pues esta lo permite. La principal forma de comprobar el desbalance de una muestra es comprobando la matriz de casos de la variable objetivo (quiebra), en este caso 14,8%.

4.2. EXPERIMENTACIÓN

4.2.1. Workflow

El *workflow* de un modelo es la manera común de denominar al conjunto de pasos por los que pasan los datos durante su construcción. El proceso seguido en este trabajo de fin de grado se puede observar de manera sintetizada en la figura 5, a continuación:

Figura 5: Workflow de implementación y creación de los modelos predictivos



Fuente: Elaboración propia

4.2.2. Implementación de la Regresión logística

Se han distinguido cuatro escenarios para la implementación de los modelos. Los tres primeros escenarios varían en cuanto al preprocesamiento de datos, pero todos utilizan la regresión logística como modelo predictivo. El cuarto caso trata la implementación de la red neuronal. En cada caso se ha procesado la muestra original de una manera diferente con el fin de obtener mejores resultados, dando lugar a una submuestra para cada escenario. Una vez se establece cuál es la submuestra para cada caso, hay que dividirla en *training set* y *test set.* El

primer conjunto de datos servirá para entrenar el modelo y que el algoritmo aprenda de los registros. El segundo pondrá a prueba los conocimientos adquiridos por el modelo, implementándolo para un nuevo conjunto de datos y contrastando la clasificación del modelo con la realidad, evaluando así su capacidad predictiva. La división en subconjuntos se realiza mediante una partición aleatoria que asigna el 80% de los datos al conjunto de entramiento y un 20% al de prueba.

En los tres primeros escenarios, tras obtener la submuestra correspondiente, se implementará la misma función de regresión logística. Para la ejecución de esta se han utilizado las funciones de la librería "sklearn". La función "LogisticRegression", es el pilar sobre el que se construye el modelo, a dicha función se le deben introducir una serie de hiperparámetros en función de cómo se quiera implementar la regresión logística. Para optimizar los hiperparámetros que se debían introducir en la función, se ha utilizado la función "GridSearchCV" para ajustar el valor de estos. De esta forma se obtiene cuál es la mejor combinación de parámetros que servirán de entrada a la regresión logística, en la función "LogisticRegression". El *solver* elegido es "newton-cg" [8] y se utiliza el parámetro "balanced" para indicar a la función que la muestra no está equilibrada, puesto que debe tenerlo en cuenta a la hora de ejecutar la regresión logística, para no caer en trampas probabilísticas. Estos parámetros son los que se implementarán para todos los casos.

- Escenario 1: Base

El escenario base toma como variables de entrada los ratios contables que contienen la información relativa a los ratios de *n* y *n-1*, es decir, "A", "A1", "B", "B1", etc., es decir, todas las variables originales. Posteriormente se suprimen los registros que contienen datos omitidos (NA), quedando la muestra con 17.381 registros.

  - Escenario 1-PCA

El escenario 1-PCA recibe la misma nomenclatura numérica que el escenario 1 o base por ejecutarse sobre la misma submuestra y con el mismo método que este. El cambio de número de escenario lleva implícito un cambio en el procesamiento de la submuestra inicial. Con el objetivo de optimizar los resultados obtenidos en el caso base, se ha llevado a cabo, sobre el mismo conjunto de datos un análisis de componentes principales (comúnmente denominado PCA de sus

[8] Este método identifica los puntos en los que la función se anula. En el contexto de la minimización, se utiliza para hallar los ceros de la derivada de una función, puesto que los puntos en los que dicha derivada se anula corresponden a un mínimo (o máximo) local de la función original.

siglas en inglés *Principal Component Analysis*). Este tipo de análisis tiene por objetivo crear variables sintéticas que recojan la misma información que las variables iniciales, de manera que con un número bajo de ellas se recoja la mayor parte de la información contenida en las variables originales.

Los resultados del análisis han determinado que se necesita una cantidad alta de variables sintéticas o componentes principales para explicar la mayor parte de la varianza y por tanto para implementar el modelo. No se va a explicar en detalle el desarrollo, pero este puede consultarse en el Anexo. Los resultados obtenidos indican que 6 variables explicarían el 57% de la varianza y 10 variables tan solo el 75% de la misma, evidenciando así la potencial pérdida de información si se implementa este método. Adicionalmente este método dificulta la comprensión del resultado obtenido, puesto que este se obtiene en función de las componentes principales, que son una composición de las variables originales que a su vez son ratios, creando una dificultad añadida a la hora de determinar cuáles son las características esenciales que determinan que una empresa va a quebrar o no, por lo que no parece el método óptimo.

- Escenario 1-SMT

En la misma línea que el escenario 1-PCA, el presente trata de mejorar los resultados obtenidos en el base. Para el desarrollo de este caso, la manera elegida para intentar hacerlo ha sido equilibrando la muestra obtenida, de manera que se cuenten con más casos de empresas quebradas, que potencialmente mejoraría la capacidad predictiva de la regresión logística. Como se ha comentado previamente, una forma de equilibrar la muestra, diferente a la utilizada al pasar el parámetro "balanced" a la función "LogisticRegression", es creando datos sintéticos que equilibren la muestra. La manera de crear estos datos es aplicando la función "smote()" que genera datos sintéticos basados en conjunto original. Tras aplicar la función se han obtenido 15.668 casos de empresas quebradas y 15.668 de empresas no quebradas, es decir, una muestra equilibrada. Posteriormente se ha aplicado la misma función de regresión logística, con los parámetros ya indicados para todos los casos. El resultado obtenido por el caso base y el obtenido en este escenario es muy similar, por lo que al no mejorar de manera sustancial los resultados no se va a profundizar en esta aproximación.

- Escenario 2

En este escenario se ha tratado de ampliar el tamaño de la muestra con respecto al caso base. La manera de intentar ampliar el tamaño muestral ha sido, como se sugirió al explorar los datos, analizando qué variables son las que tienen más valores perdidos y menor significancia y suprimirlas antes de la eliminación de las filas con algún valor perdido. De esta forma se ha prescindido de las variables que hacen referencia a los ratios: "D", "D1", "H", "H1", "M","M1", "T", "T1", "P"," P1", "Q" y" Q1". Posteriormente se han suprimido las columnas que muestran multicolinealidad o alta correlación entre sí, considerando una correlación mayor a 0,85 como alta. Por último, se han omitido los registros que tenían valores perdidos, obteniendo así una muestra que cuenta con 20.038 registros para implementar la función de regresión logística, esto es 2.257 registros más que en el caso base. Posteriormente, se ha implementado la función "LogisticRegression" con los parámetros ya mencionados. Con estos registros adicionales, se ha producido una mejora en los resultados.

- Escenario 3

El tercer escenario quiere comprobar el desempeño de las variables sintéticas creadas que en principio reducían la multicolinealidad de las variables originales, en el caso de un ratio de un año y del anterior. Para ello se han utilizado como entrada las variables con la información relativa al año *n-1* y el ratio (*n/n-1*), es decir "A1", "AR", "B1", "BR", etc. Después se han eliminado los registros que contaban con valores omitidos, obteniendo 17.381 registros de muestra, siendo esta del mismo tamaño que en el caso base. Por último, se han querido eliminar las variables que aún así mostraban una alta colinealidad para tratar de obtener un modelo más preciso, puesto que, en principio, la multicolinealidad dificulta la obtención de modelos robustos, como ya se ha explicado en el epígrafe 4.1. Una vez realizados todos los pasos, se ha implementado la función "LogisticRegression" con los mismos parámetros utilizados hasta ahora en el resto de casos. Con este escenario se han logrado unos resultados similares a los obtenidos para el segundo escenario.

4.2.2.1. Importancia de las variables

La regresión logística es un método que permite explicar qué ha llevado al modelo a tomar las decisiones clasificatorias y por tanto dar transparencia al proceso. La demanda de información acerca del proceso de toma de decisión es necesaria para el objetivo que trata este trabajo, ya que las empresas de M&A quieren comprender el porqué de la decisión del modelo para poder analizar más en profundidad esos factores, tratando de evitar en la mayor medida posible las "cajas negras". Esto además de dar confianza en que el modelo puede explicar el porqué está dando ese

resultado en función de las variables de negocio, también ofrece una valiosa información de cuáles son las variables más significativas a la hora de realizar una clasificación, y en qué sentido actúa cada variable respecto a la clasificación. En el caso de la regresión logística, si las variables de entrada están normalizadas, la manera de explicarlo es mediante el valor relativo de los coeficientes de la regresión. Estos indicarán la importancia relativa de cada variable y en qué sentido de la clasificación aportan valor, bien positivo aumentando la probabilidad de ser clasificado como quebrado, bien negativo aumentando la probabilidad de ser clasificado como no quebrado. Para ello, se ha realizado una función que asigna la importancia relativa a cada variable usando los coeficientes de la regresión con los datos normalizados, mostrándolo en un gráfico de barras, como se puede ver en los resultados obtenidos en las figuras 8,9 y 10 que se muestran en el siguiente epígrafe.

El signo de la importancia indica si la probabilidad de quiebra aumenta o disminuye con el incremento del valor absoluto de la variable. Es decir, si hay una variable con un alto negativo, significa que, si la empresa tiene un valor alto en ella, es más probable que sea clasificada como no quiebra, al menos si solo se tiene en cuenta esa variable. Al contrario, un alto valor positivo de la importancia de una variable indicará que cuanto mayor, más probabilidad tendrá de ser clasificada como quebrada. Al interpretar las variables más relevantes, hay que tener en cuenta que al haber variables iniciales que están muy correlacionadas, se puede trasladar la importancia relativa a ellas también. Es decir, si C y D están muy correlacionadas y se ha eliminado D para hacer el modelo y C sale como variable muy relevante, también lo es D, aunque no aparezca en la lista de importancia de las variables. Es más, al eliminar variables muy correlacionadas para evitar la colinealidad de ellas, dependiendo del orden de los pasos del preprocesado, se puede eliminar la una o la otra y que no entre en modelo como variable de entrada.

4.2.3. Implementación de la Red Neuronal

- Escenario 4

Para el cuarto escenario, se ha usado una red neuronal con las variables *n* y *n-1*, ya que este tipo de modelo es capaz de realizar en sus capas de neuronas intermedias los cálculos o *features* que sean más adecuados. Las variables han sido escaladas previamente, para facilitar la implementación de la red neuronal, como dice la literatura. Para las entradas se han eliminados los valores perdidos y se han optimizado los hiper parámetros de la red neuronal y diseñado la misma con la libreria “keras”, usando como motor “Tensorflow”.

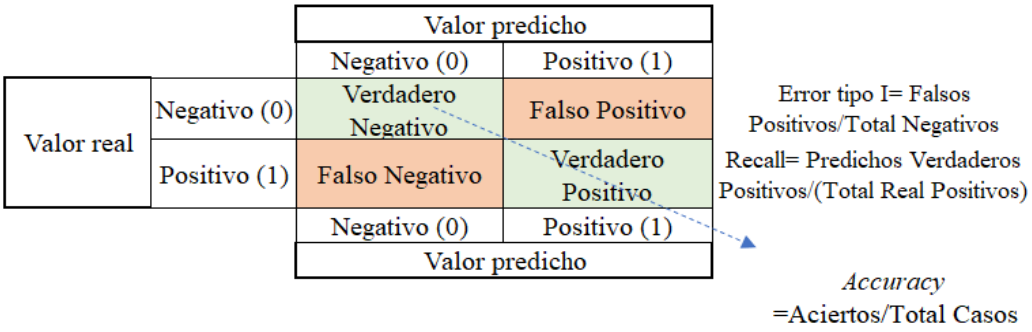
4.3. RESULTADOS GENERALES

4.3.1. Matrices de confusión

La matriz de confusión es un elemento esencial para medir el desempeño de un modelo de *machine learning*. Dentro de ella se encuentra la cantidad de empresas que el modelo ha predicho como quebradas y no quebradas y cuál es su clasificación real de la forma que se muestra en la figura 6, lo que da una visión interesante para medir el tipo de error que comete el modelo.

En primer lugar, el error tipo I hace alusión a los falsos positivos, en otras palabras, clasificar como quebrada una empresa sana y el error tipo II se refiere a los falsos negativos, es decir, a calificar como sana una empresa quebrada. Un alto porcentaje de error tipo I podría ser tenido en cuenta como un aspecto negativo, pues podría derivar en mala reputación para una empresa que no ha quebrado. Sin embargo, yendo más allá el error de tipo I puede corresponder con empresas que están siendo artificialmente sostenidas, que deberían haber quebrado o que simplemente están posponiendo la decisión de declararse oficialmente en concurso. Por ello, el error tipo I es una medida muy relevante para el propósito de este estudio, pues se trata de un modelo de uso interno para empresas de M&A que en todo caso realizarán un estudio posterior de viabilidad sobre las entidades que el modelo ha calificado como quebradas, sin *a priori* desvelar información que derivaría en mala reputación. Esta métrica es muy relevante para el análisis realizado en este trabajo de fin de grado por el objetivo que persigue, y no pretende minimizar este tipo de error, a diferencia de lo que otros autores han propuesto en sus estudios, que buscan minimizarlo, como Beaver (1966). El motivo de no minimizar el error tipo I es que a las empresas en el marco de M&A les suscita especial interés conocer qué entidades podrían tener características similares a las quebradas para así, tras un análisis posterior, confirmar y maximizar su oportunidad de negocio.

Figura 6: Ejemplo de matriz de confusión con métricas relevantes



Fuente: Elaboración propia

La figura 6, muestra un ejemplo del contenido de una matriz de confusión, de donde se pueden extraer diferentes métricas como son la especificidad, la sensibilidad o la precisión entre otras. Para el propósito de la investigación llevada a cabo en este trabajo y conforme al objetivo de maximizar la capacidad de negociación de las empresas de M&A las medidas más relevantes para el caso se encuentran resumidas en la figura 7 y la tabla 5. Las medidas en las que se va a poner énfasis son el *accuracy*, el *recall* y la tasa de falsos positivos.

En segundo lugar, el *accuracy* hace referencia a la precisión total del modelo, y sigue la fórmula:

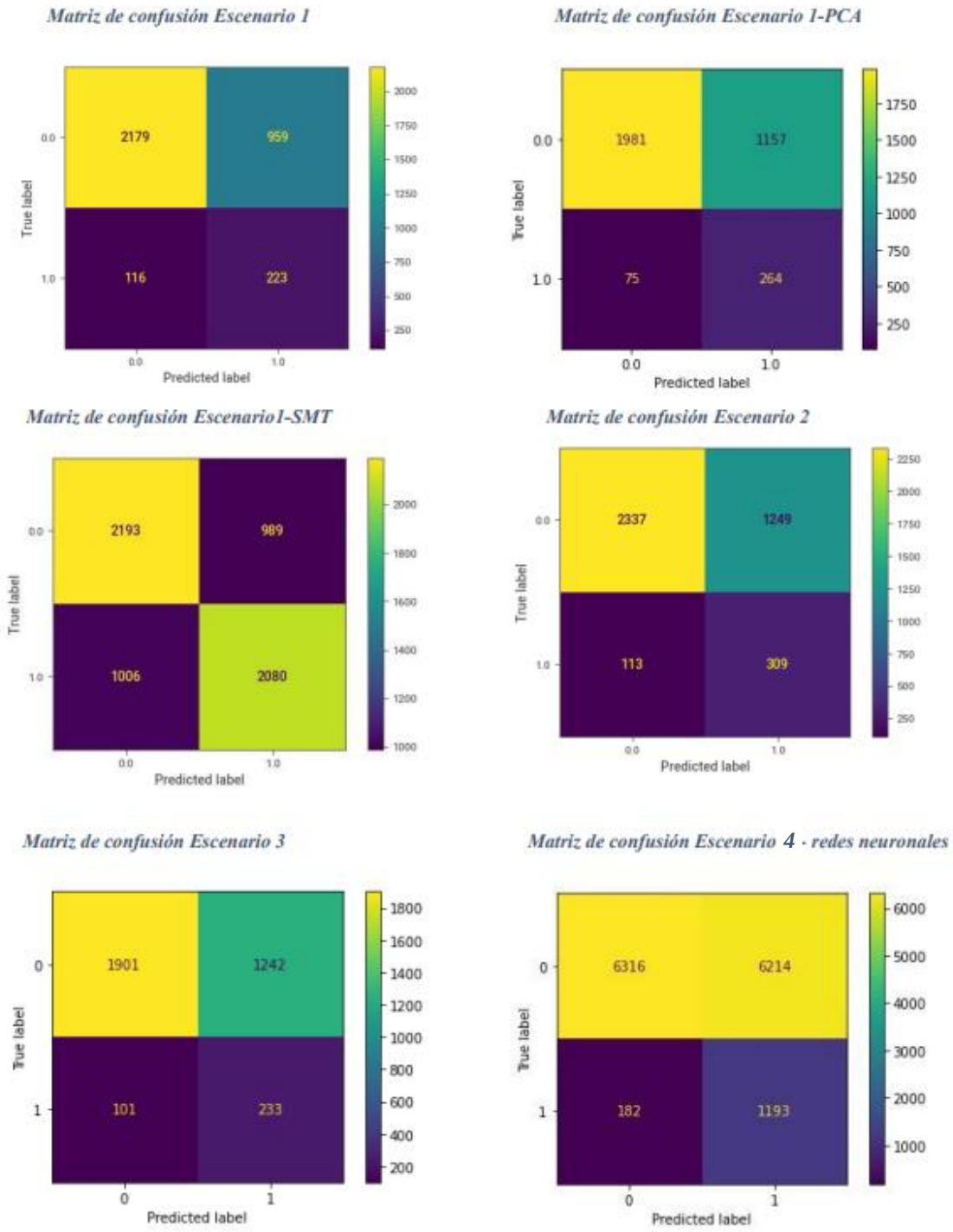
$$Accuracy = \frac{Verdaderos\ positivos + Verdaderos\ negativos}{Total\ registros}$$

Esta métrica mide la tasa de acierto tanto de empresas quebradas, como de no quebradas sobre el total de empresas puestas a prueba (el conjunto de test). *A priori*, interesa tener un *accuracy* alto, puesto que el objetivo de los modelos de *machine learning*, en general, es predecir correctamente. Sin embargo, esta medida no es la más relevante, puesto que como se explicó cuando se hizo alusión a la necesidad de equilibrar la muestra, se podrían obtener modelos con un *accuracy* muy alto únicamente basados en la probabilidad, que no aportarían información acerca de las empresas. Es así como a pesar de que se deba tener en cuenta el *accuracy* a la hora de decidir qué modelo es mejor, no va a ser la métrica más relevante.

En último lugar, la métrica de *recall* es indicativa de la tasa de acierto en la predicción de quiebra sobre el total de quebrados reales. Esta medida es mucho más interesante que el *accuracy*, puesto que trata de medir con qué probabilidad el modelo acierta cuando predice que una empresa quebrará.

En la figura 7 se encuentra el resultado de las matrices de confusión obtenidas de la ejecución de los diferentes escenarios y modelos.

Figura 7: Matrices de confusión



Fuente: Elaboración propia

La tabla 5 muestra una comparativa a modo resumen de los resultados obtenidos por los diferentes escenarios y modelos. De los resultados que se muestran en esta tabla, se puede observar que el escenario 4 es el que mejor resultado ha obtenido en términos de *recall*, seguido del escenario 1-PCA y del escenario 2.

El escenario 1-PCA, ha sido desechado como la mejor alternativa porque al haber implementado el análisis de componentes principales se dificulta la explicabilidad de las variables. Por tanto, de entre los modelos que se han llevado a cabo utilizando regresión logística, el segundo escenario es el que mejor *recall* obtiene. Siendo esto así, se podría decir que el segundo escenario es el que mejor predice la quiebra.

Como para implementar la regresión logística en el escenario 2 se han utilizado más registros reales que para el resto de casos de aplicación de regresión logística, se puede atribuir la mejora en el *recall* con respecto a los escenarios 1 y 3 visible en la tabla 4, al incremento del tamaño muestral, puesto que la función implementada es la misma. Con esto se puede concluir que, en efecto, las variables eliminadas para llevar a cabo el escenario 2, eran de poca importancia y contenían muchos valores perdidos, puesto que predicen igual, o mejor que el resto de los modelos. Dicho esto, los modelos que según el *recall* son más interesantes, son el 2 y el 4, aunque de estos el único que permitirá explicar la importancia de las variables será el modelo 2, por el carácter heurístico de las redes neuronales. Es cierto que existen técnicas como la SHAP para determinar la importancia de las variables de los modelos basados en redes neuronales. Sin embargo, la aproximación llevada a cabo en este estudio para elegir los métodos a utilizar no contempla el uso de técnicas cuya implementación no consta en la literatura consultada sobre el tema.

Tabla 5: Tabla resumen de resultados obtenidos

	Accuracy	***Recall***	**Error Tipo I**
Escenario 1	0,69	0,66	0,31
Escenario 1-PCA	0,65	0,78	0,37
Escenario 1-SMT	0,68	0,67	0,31
Escenario 2	0,66	0,73	0,35
Escenario 3	0,61	0,70	0,40
Escenario 4	0,54	0,87	0,50

Fuente: Elaboración propia

Dentro del *accuracy*, hay que destacar que por lo general todos los modelos implementados tienen un *accuracy* bajo. Es interesante que, a pesar de que *a priori* por la complejidad del modelo y la tipología de los datos parecía que las redes neuronales iban a ser las que mejor desempeño obtuvieran, no ha sido así, ya que son las que peor puntuación han obtenido. Dentro de los modelos de regresión logística el escenario 1 y el 1-SMT han sido los que mejor puntuación han obtenido, sin embargo, con un *recall* más bajo que el escenario 2. De esto se puede deducir que la mejora predictiva de los escenarios 1 y 1-SMT, que hace que aumente el *accuracy*, es en términos de empresas sanas, que es el más común, y al tener un *recall* más bajo no se puede decir que sean mejores en términos de predicción que el modelo 2.

Poniendo el foco ahora en el error de tipo I haciendo referencia a los datos presentados en la tabla 4, se puede observar que el escenario 2 tiene significativamente un menor porcentaje de error que el escenario 4, 0,35 frente a 0,5 respectivamente. Esta medida es indicativa de que en el cuarto escenario se sobre predice la quiebra. Sin embargo, al no poder conocer las razones que hay detrás de ello, por su carácter de algoritmo "caja negra", no se puede explicar la relación que existe entre los datos y esta predicción excesiva; que podría deberse a la mera probabilidad o a motivos más complejos. El valor de 0,35 del error de tipo I obtenido en el segundo escenario, puede ser un factor interesante por comentar, ya que es sugerente entender qué variables son las que explican el modelo para poder sacar más conclusiones. Este estudio exhaustivo se realizará en el siguiente apartado.

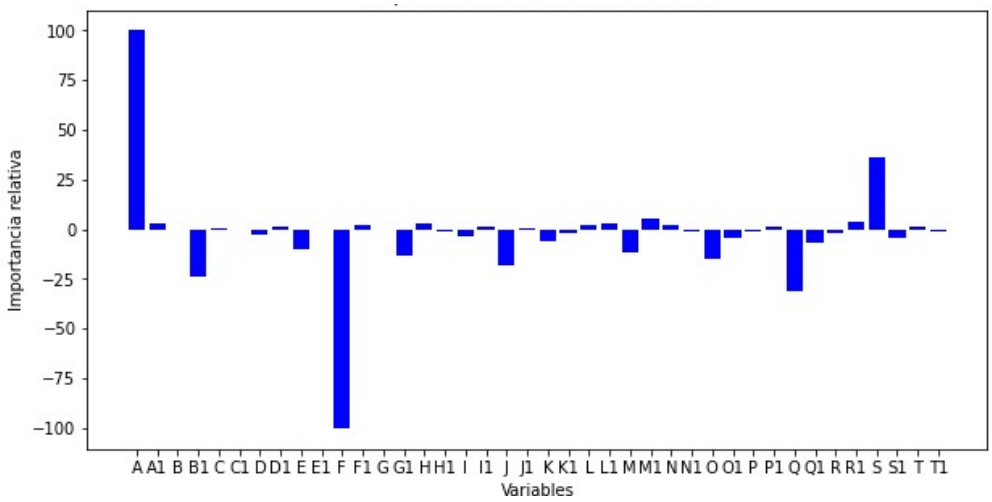
De estos resultados llama la atención el bajo desempeño del tercer escenario, que *a priori* pretendía disipar la multicolinealidad entre las variables, creando una sintética nueva correspondiente a la variación. Dicho esto, se puede concluir que las variables sintéticas no han aportado un diferencial al modelo, ya que es más importante la situación de la empresa que la variación del ratio. Es decir, que lo realmente relevante es la situación empresarial, no los cambios que han sucedido en ella de un año a otro.

4.3.2. Importancia de las variables

El resultado gráfico de la importancia de los escenarios primero, segundo y tercero se encuentra en las figuras 8, 9 y 10 respectivamente. Para analizar correctamente los diagramas de barras de la importancia relativa de las variables conviene recordar la explicación que existe detrás de la misma. En este sentido, una variable que obtenga un resultado positivo elevado en términos de relevancia debe ser entendida de forma que, si una empresa tiene un valor muy alto en la misma en términos absolutos, aumenta su probabilidad de ser calificada como quebrada por el modelo. En sentido contrario, si una empresa tiene un valor alto en una variable que tiene una importancia

relativa muy negativa, su probabilidad de ser clasificada como quebrada por el modelo disminuye, aumentando su probabilidad de ser clasificada como empresa sana.

Figura 8: Importancia de las variables Escenario 1-Base



Fuente: Elaboración propia

En la figura anterior se muestra la importancia relativa de las variables del escenario base. De ella se puede concluir que los ratios más relevantes que aumentan la probabilidad de ser calificada como quebrada son A y S. Mientras que las variables F, Q y en menor medida B1 entre otras, ayudan a predecir la no quiebra.

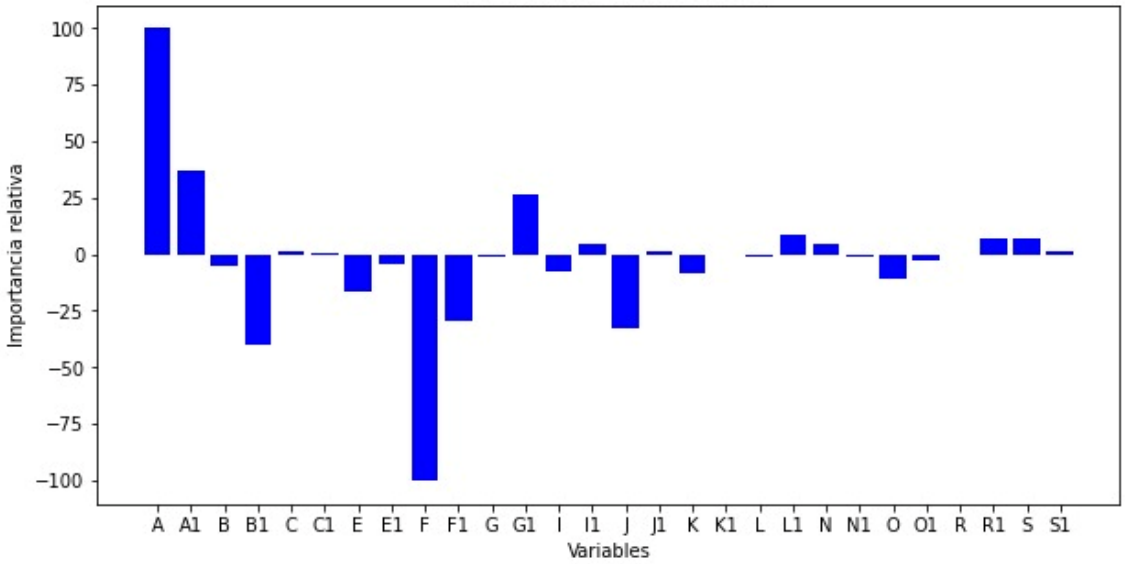
Con esto se puede comprobar que para el escenario 1 lo más relevante es tener un buen ratio A, es decir, un ratio positivo alto de ingresos netos sobre total de activos. Una posible explicación encajaría con empresas pequeñas que carecen de activos, pero no deja de ser algo sorprendente y debería ser estudiado con más profundidad, al igual que el ratio S, de fondo de maniobra sobre ingresos de ventas del año en curso, que implica facilidad de pagos y liquidez pero que el modelo pondera para predecir quiebra. Esto, junto a su bajo desempeño, hace que este modelo no sea el seleccionado y se prefieran otras aproximaciones.

Por otro lado, como variables explicativas que ayudan a predecir la no quiebra se observan altos indicadores del ratio F, ingresos de ventas sobre activos del año en curso, lo que supone un buen desempeño de la compañía. Es decir, que con pocos activos se consiguen ingresos por ventas altos, lo que podrían estar relacionado con una alta rentabilidad sobre los activos. Es interesante observar también que el año anterior no tiene efecto para este ratio. Otro ratio que indica una menor probabilidad de quiebra es Q, de existencias entre ingresos por ventas. Es sorprendente que un alto valor en este ratio aumente la probabilidad de ser clasificado como no quebrado, puesto que lo habitual es intentar que este sea lo menor posible, minimizando las existencias, y en este caso no es así. Podría indicar que, en el sector de la restauración, tener un unas altas

existencias es importante para ofrecer un servicio de calidad, quizás por permitir una variada oferta o por no perder ventas por roturas de stock. También podría justificarse si la materia prima supone una parte pequeña del total de los gastos, es decir que compensase el coste de las existencias frente al coste de oportunidad de las ventas por tener o no tener esas existencias para la venta. En cualquier caso, habría que investigar en profundidad las causas para verificar las hipótesis y encontrar la causa real.

Por último, la siguiente variable por relevancia, aunque en menor medida es el ratio B1, de activo corriente (*n-1*) entre pasivo corriente (*n-1*), indicador de solvencia. Esta métrica es razonable, ya que habitualmente una empresa solvente no quebrará. Por tanto, cuanto mayor solvencia, menor es la probabilidad de quiebra.

Figura 9: Importancia de las variables Escenario 2



Fuente: Elaboración propia

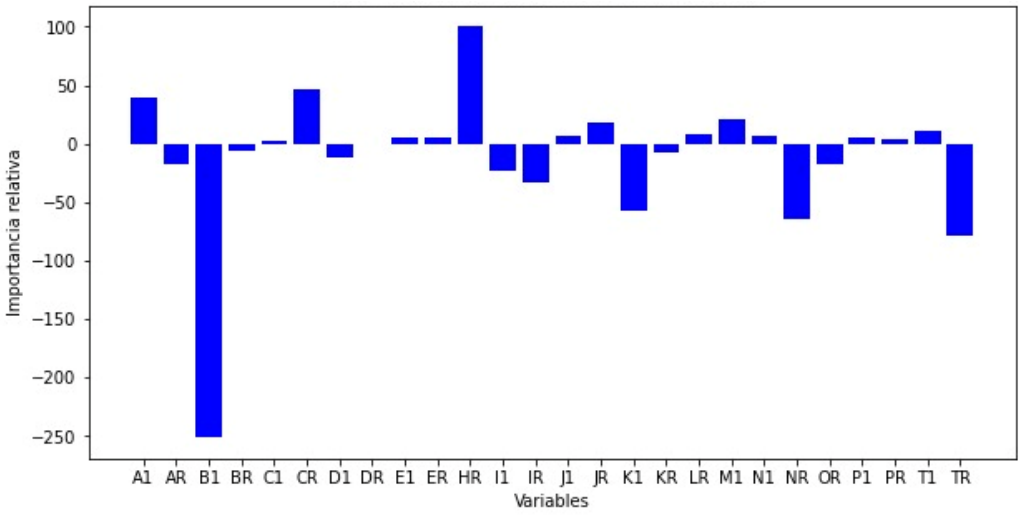
Como se puede observar en la figura 9, para el escenario 2, las variables que aumentan la probabilidad de quiebra en mayor medida son los ratios A, A1 y G1. Por otro lado, los ratios que hacen que disminuya la probabilidad de quiebra, es decir afectan en sentido negativo a la quiebra son el F, B1, J y F1.

En el segundo escenario, se contaba con mayor cantidad de registros que en el caso base por la estrategia seguida a la hora de eliminar datos perdidos. Según este modelo el ratio A aumenta la probabilidad de quiebra, siendo este ingresos netos sobre total activos del año en curso; también aumenta la probabilidad de quiebra, aunque en menor medida, el mismo dato del año anterior, recogido en la variable A1. Se entiende que la explicación es la misma que en el escenario base y que al ser este resultado contraintuitivo se deberá estudiar en profundidad.

También aumenta la probabilidad de quiebra cuanto mayor sea el ratio G1, que hace referencia al activo corriente menos existencias entre pasivo corriente. Un valor alto en este ratio indicaría que la empresa tenía muy pocas existencias en relación con su activo corriente. Como se ha indicado previamente, habría que estudiar el tamaño de la entidad, puesto que, que una compañía de pequeño tamaño cuente con pocas existencias es acorde a su demanda, sin embargo, que así lo haga una empresa de gran tamaño, dificulta su funcionamiento.

En los que respecta a las variables que aumentan la probabilidad de no quiebra, el escenario 2 incluye el ratio F, que ya ha sido explicado para el escenario base. La siguiente variable en importancia es el ratio B1, activo corriente entre pasivo corriente del año anterior, es decir, liquidez, ya que cuanto mayor liquidez, más dificultad de quiebra. Sin embargo, un valor alto de este ratio puede indicar otras carencias, como coste de oportunidad para inversiones. Se entiende que en el sector de la restauración este factor es muy importante ya que cobran a la mayoría de sus clientes al contado, mismo proceder que utilizan para el pago a sus proveedores. El siguiente ratio en importancia para predecir no quiebra es J, ingresos netos entre patrimonio neto, lo que implica poca inversión de los socios para el volumen de ventas de la empresa, aunque también pocas reservas, lo que podría leerse en el sentido de tener una alta capacidad de generar ingresos con poco patrimonio de la empresa.

Figura 10: Importancia de las variables Escenario 3



Fuente: Elaboración propia

Para el tercer escenario las variables positivas más determinantes de quiebra son HR, CR y A1, mientras que B1, TR, NR y K son las que predicen en sentido negativo, es decir no quebrar como así se indica en la figura 10.

Hay que recordar que, en el tercer escenario se han incluido los ratios del año anterior y un nuevo ratio de variable año en curso entre la variable del año anterior. Uno de los ratios que más importancia tiene para aumentar la posibilidad de quiebra es HR, que es la variación relativa de total pasivo entre total activo. El principal motivo de esto puede ser un aumento del pasivo o una disminución del patrimonio neto, lo que indicaría en ambos casos que ha aumentado la financiación externa, disminuyendo la situación de liquidez y favoreciendo la probabilidad de quiebra. El siguiente ratio en importancia para aumentar la probabilidad de quiebra es CR, que representa el fondo de maniobra entre total activo. Esto en general puede parecer positivo asumiendo que aumente el fondo de maniobra, aunque la causa puede ser también la disminución del total del activo. Por último, con menor relevancia el modelo tres incluye A1, ingresos netos entre total activo del año anterior, que ya ha sido comentado en los casos anteriores.

Como variables que aumentan la probabilidad de ser clasificado como no quebrado en el caso tres, se encuentran B1, TR, NR y K. La explicación por la que el ratio B1, activo corriente entre pasivo corriente del año anterior tiene importancia ya se ha explicado en anteriores escenarios. La segunda variable en relevancia es TR, que es el ratio pasivo fijo (*n*) / total activo (*n*) respecto al año anterior, es decir la variación interanual del ratio; indicaría que tener un incremento de la financiación a largo plazo respecto al total de activo en el periodo, este factor es relevante para que se prediga no quiebra. La siguiente variable en importancia es NR, que es el ratio de pasivo corriente (*n*) / total activo (*n*) respecto al año anterior; es extraño que tenga efecto en el mismo sentido que la variable NR y se deberían estudiar con más detalle los posibles motivos. Por último, un alto valor de K, que representa tesorería (*n*) / total activo (*n*), explicaría mayor dificultad para quebrar, lo que es muy razonable puesto que supone tener facilidad de pagos a corto plazo.

De la tabla 6 se puede extraer una visión general que resume las variables determinantes para la quiebra y la no quiebra para cada modelo, concluyendo que son diferentes a pesar de que *a priori* deberían ser similares. Por ello, es importante comentar que en el *dataset* hay mucha colinealidad, debida a que diferentes ratios dependen de las mismas variables o están relacionados indirectamente por su composición. Teniendo entonces en cuenta que, cada modelo elimina unas variables para disminuir el efecto de la colinealidad y no lo hace siempre en el mismo orden o momento del procesamiento; es preciso mencionar que estos cambios hacen que se eliminen variables diversas.

Tabla 6: Comparativa de la importancia de las variables en cada escenario

	Variables determinantes de quiebra	Variables determinantes de NO quiebra
Escenario 1	A, S	F, Q, B1
Escenario 2	A, A1, G1	F, B1, J, F1
Escenario 3	HR, CR	B1, TR, NR, K

Fuente: Elaboración propia

Analizando las variables que más influyen en los tres escenarios para aumentar la probabilidad de quiebra se observa que en los tres coinciden en el ratio A, ingresos netos entre total activo. Sin embargo, no lo hacen en el resto de ratios, S, fondo de maniobra entre ingreso por ventas (caso A); G, activo corriente menos existencias entre pasivo corriente (caso dos) y HR Pasivo total entre activo total, CR, fondo de maniobra entre total activo (caso tres). Aparentemente todas las variables no coincidentes tienen en común el efecto del activo corriente como parte importante del ratio. Sin embargo, al no guardar una correlación relevante entre ellas, la variación en torno a la relevancia de los ratios puede deberse al efecto indirecto del activo corriente o a las peculiaridades de cada modelo, que derivan en diferentes tasas de acierto o *accuracy* como se ha comentado en el apartado anterior.

Por otro lado, analizando qué variables influyen para aumentar la probabilidad de no quiebra, se observa que los tres modelos coinciden en que la variable B1, correspondiente al Activo Corriente (*n-1*) / Pasivo Corriente (*n-1*), es determinante. Que los escenarios coincidan en este ratio tiene sentido debido a que la solvencia de una empresa es esencial para determinar su situación de liquidez; y por tanto que una empresa sea muy solvente, la califica como menos propensa a ser ilíquida. Por otro lado, los escenarios 1 y 2 coinciden en que el ratio F, que corresponde a ingresos por ventas (*n*) / total activo (*n*), también es determinante. Este ratio indica poca inversión en relación con las ventas, lo que supone un buen desempeño de la compañía, ya que con pocos activos se consiguen altas ventas, que podría estar relacionado con alta rentabilidad sobre los activos. Las variables Q, J, TR, NR, K aparentemente no guardan relación, más allá de que son dependientes bien de los ingresos por ventas o del total activo, por lo que se puede concluir que estos dos factores tienen un efecto directo importante sobre la predicción de no quiebra de los modelos. También es preciso recordar que la variación de importancia de las variables puede deberse a como se ha explicado previamente, las iteraciones del modelo o el orden de eliminación de variables correlacionadas y no solo a la información contenida en ellas.

CAPÍTULO 5. CONCLUSIONES

Los resultados obtenidos tras la implementación de técnicas predictivas basadas en la estadística -regresión logística- y en la inteligencia artificial -redes neuronales-, para clasificar las empresas de parte del sector de la restauración en España dejan margen de mejora. Los modelos construidos que utilizan la regresión logística han obtenido una precisión (*accuracy*) del 0,658% en media, yendo desde el 0,61% al 0,69%. Sin embargo, el modelo implementado con la base de redes neuronales tan solo ha obtenido un 0,54% de precisión. Si se compara con otros trabajos realizados en el mismo ámbito, los resultados son algo más bajos de lo esperado.

Es importante recordar el objetivo que este trabajo trataba de conseguir, pues al tratar de buscar una herramienta que sirva a las empresas de M&A para identificar potenciales sociedades que adquirir, la precisión no es la métrica más relevante, si bien es necesario comentarla. En este sentido, lo más interesante no es tener un alto porcentaje de acierto en todos los casos de la predicción, como por ejemplo los falsos positivos, si no que el acierto sea muy alto para la casuística más relevante en el caso concreto de M&A. Es así como la estrategia perseguida por este tipo de entidades es la de detectar empresas que van a quebrar y también encontrar otros patrones, como las sociedades que probablemente deberían haberlo hecho por sus características y, sin embargo, no han quebrado, por lo que el modelo no acertaría, si bien aportaría información relevante.

Los modelos obtenidos son razonables en líneas generales, si bien el modelo implementado en el escenario 2 proporciona una tasa más alta de errores tipo I, con un *accuracy* similar al del resto de escenarios. Este escenario trata de mejorar el caso base mediante la eliminación de las variables que más valores perdidos contenían y menos relevantes eran, previo a la eliminación del resto de registros omitidos, de manera que se aumenta el número de registros de la submuestra. Si además se tiene en cuenta la tasa de *recall*, se puede concluir que en efecto predice las entidades quebradas mejor que el resto de modelos y, por tanto, las empresas mal clasificadas como quebradas serán potencialmente más interesantes para las entidades adquirentes. Además, parte del atractivo de este modelo reside en la explicación aportada, es decir, la relevancia que se le otorga a los ratios del modelo.

Como ha quedado latente al comparar los resultados obtenidos en los escenarios 1-SMT y 2, aumentar la muestra deriva en mejores resultados. De esta forma, sería muy interesante aumentar el alcance de los datos, bien con otra base de datos que permita más descargas, bien mediante una depuración del set de datos diferente a la empleada. Incluso se podría pensar en una estrategia basada en emplear los ratios que menos valores perdidos tengan. También podría

realizarse el estudio en un sector que contuviera menos valores perdidos, de manera que se podría comparar la implementación de las mismas técnicas en diferentes sectores, pudiendo no solo comparar el efecto del tamaño muestral, sino también de otros factores que son característicos de cada sector empresarial.

Otro aspecto por destacar es que no hay que confiar en los criterios contables estándar para intuir dificultades financieras de una empresa. Se ha visto en el trabajo que algunos de los ratios que el modelo ha marcado como más relevantes para predecir quiebra o la no quiebra no son los que intuitivamente se pensarían. En ocasiones, como en el ratio ingresos netos/total activo se ha visto que incluso afectan en sentido contrario al previsto, ayudando a predecir la quiebra, cuando la lógica más directa hace pensar lo contrario de lo que indican. Este aspecto, debe ser estudiado con más profundidad para determinar su causa y poder sacar conclusiones más relevantes, tanto si se debe a peculiaridades de los modelos analíticos, como si es por la idiosincrasia del sector.

También es importante comentar que los ratios que los modelos indican como determinantes para la predicción no son los mismos en todos los escenarios, variando así, en función del preprocesamiento de los datos. En el estudio con los diferentes escenarios se obtienen algunos ratios comunes y otros exclusivos de un escenario. Parte de la causa es que muchos de los ratios empleados tiene una alta colinealidad y, aun siendo ratios distintos, reflejan casi la misma información predictiva de probabilidad de quiebra. Así, se debe entender esto en la perspectiva de la extensa literatura sobre el tema de predicción de quiebras donde se proponen multitud de ratios siendo algunos recurrentes y otros intercambiables entre sí con una escasa variación en sus resultados. Como línea de investigación futura y atendiendo a las cuestiones que se han suscitado durante el desarrollo de este trabajo, propongo realizar un estudio de la predicción de quiebras basándolo en la información de las partidas contables como tal y no a través de ratios, para comparar su desempeño.

Respecto al modelo de red neuronal, aunque en la actualidad tienen una excelente reputación, sobre todo últimamente con los modelos de inteligencia artificial generativa y de tratamiento de lenguaje natural con ChatGPT, han tenido un resultado discreto. Entre los posibles motivos que en este estudio pueden haber influido en la escasa mejora respecto a regresión logística, mencionar el bajo número de registros de la muestra y que las variables de entrada ya eran ratios elaborados, cuando una red neuronal funciona mejor con información menos elaborada.

Con todo, sería muy interesante realizar un análisis similar al llevado a cabo aumentando los registros, e incluso aportando información no contable, que según la literatura mejora el desempeño de los modelos. Si bien, encontrar este tipo de información puede ser complicado, teniendo en cuenta las limitaciones suscitadas con respecto a los datos disponibles contables, que *a priori* son más sencillos de conseguir.

También sería interesante como línea de investigación complementaria, llevar a cabo un análisis de las tendencias en M&A en el sector de la restauración. Con este se podría comprender cuáles son los factores decisivos para las adquisiciones, más allá del presupuesto bajo precio por activos valiosos.

CAPÍTULO 6. BIBLIOGRAFÍA

AARONALLEN & associates global restaurant consultants. (5 de Octubre de 2022). *Fusiones y Adquisiciones de Restaurantes: Reconfigurando la Industria*. Recuperado el 25 de Abril de 2023, de https://aaronallen.com/blog/espanol/fusiones-y-adquisiciones-de-restaurantes

Agarwal, A. (1993). *Neural networks and their extensions for business decision-making.* Ohio: Ohio State University.

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework tool selection. *Expert Systems With Applications (Elsevier)*(94), 164-184.

Altman, E. I. (1968). Financial ratios discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*(23(4)), 589-609.

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the U.S. market. *Abacus*(43(3)), 332-357.

Anandarajan, M., Anandarajan, A., & Lee, P. (2001). Bankruptcy Prediction of Financially Stressed Firms: An Examination of the Predictive Accuracy of Artificial Neural Networks. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 69-81.

Argadoña, A. (1998). *La teoría de los stakeholders y el bien común.* Universidad de Navarra, División de Investigación. Barcelona: IESE. Obtenido de https://media.iese.edu/research/pdfs/DI-0355.pdf

Aziz, M. A., & Dar, H. A. (2006). Predicting corporate bankruptcy: where we stand. *Corporate Governance*(6), 18-33.

Back, B., Laitinen, T., Sere, K., & van Wezel, M. (1996). Choosing Bankruptcy Predictors Using Discriminant Analysis, Logit Analysis, and Genetic Algorithms. *Turku Centre for Computer Science, Technical Report*(40), 1-18. Obtenido de https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=61522e036f34e5cbae416b87efd7356879a1b3a6

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research, 4*, 71-111. doi:https://doi.org/10.2307/2490171

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A Review of Bankruptcy Prediction Studies: 1930 to Present. *Journal of Financial Education, 33*, 1-42. Obtenido de https://epublications.marquette.edu/cgi/viewcontent.cgi?article=1025&context=account_fac

Collins, R. A. (1980). An Empirical Comparison of Bankruptcy Prediction Models. *Financial Management, 9*(2), 52-57. doi:https://doi.org/10.2307/3665168

Comisión de las Naciones Unidas para el Derecho Mercantil Internacional. (2006). *Guía Legislativa sobre el Régimen de la Insolvencia.* Nueva York: Organización de las Naciones Unidas. Recuperado el 20 de febrero de 2023, de https://uncitral.un.org/sites/uncitral.un.org/files/media-documents/uncitral/es/05-80725_ebook.pdf

Insituto Nacional de Estadística. (2022). *Notas explicativas CNAE-2009.* INE. Recuperado el 3 de Marzo de 2023, de https://www.ine.es/daco/daco42/clasificaciones/cnae09/notasex_cnae_09.pdf

Instituto Nacional de Estadística. (10 de Noviembre de 2022). *Demografía armonizada de empresas año 2020.* INE. Obtenido de https://www.ine.es/prensa/dae_2020.pdf

Iturriaga, F. L., & Sanz, I. P. (2015). Bankruptcy visualization and prediction using neural networks: A study of US commercial banks. *Expert Systems with Applications*(42(6)), 2857-2869.

Jefatura de Estado. (06 de Septiembre de 2022). Ley 16/2022, de 5 de septiembre, de reforma del texto refundido de la Ley Concursal, aprobado por el Real Decreto Legislativo 1/2020, de 5 de mayo, para la transposición de la Directiva (UE) 2019/1023 del Parlamento Europeo y del Consejo, de 20 de junio d. *Boletin Oficial del Estado*(214). doi:https://www.boe.es/eli/es/l/2022/09/05/16/con

Jeng, B., Jeng, Y. M., & Liang, T. P. (1997). FILM: A fuzzy inductive learning method for automated knowledge acquisition. *Decision Support Systems*(21(2)), 61-73.

Jones, S., & Hensher, D. A. (2008). *Advances in credit risk modelling and corporate bankruptcy prediction.* New York: Cambridge University Press.

Kwon, Y., & Cho, Y. (2016). The prediction of corporate bankruptcy using decision tree and ensemble methods: The case of Korea. *Expert Systems with Applications*(44), 36-43.

Laitinen, E. K., & Laitinen, T. (2000). Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International Review of Financial Analysis, 9*(4), 327-438. doi:https://doi.org/10.1016/S1057-5219(00)00039-9

Minbiole, E. N., Poli, P. M., & Haka, S. F. (2015). The Evolution of Accounting. *Journal of Accountancy*(220(3)), 30-36.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankrupcy. *Journal of Accounting Research, 18*, 109-131.

O'Leary, D. E. (21 de Diciembre de 1998). Using neural networks to predict corporate failure. *Intelligent Systems in Accounting, Finance and Management, 7*, 187-197. Obtenido de https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291099-1174%28199809%297%3A3%3C187%3A%3AAID-ISAF144%3E3.0.CO%3B2-7

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*(1), 81-106.

Raj, A. (30 de Marzo de 2022). *Towards Data Science*. Obtenido de https://towardsdatascience.com/everything-about-svm-classification-above-and-beyond-cc665bfd993e

Real Academia Española. (20 de febrero de 2023). *Diccionario de la lengua española*. Obtenido de https://dle.rae.es/quiebra

Rodríguez, M., Piñeiro, C., & de Llano, P. (s.f.). *Predicción de insolvencia y fracaso financiero: medio siglo después de Beaver(1966). Avances y nuevos resultados.* FISYG. A Coruña: Universidad de A Coruña.

Shumway, T. (Enero de 2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business, 74*(1), 101-124. doi:https://doi.org/10.1086/209665

Sun, J., Li, H., Huang, Q.-H., & He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling and featuring approaches. *Knowledge-Based Systems, 57*, 41-56.

Tascón Fernández, M. T., & Castaño Gutierrez, 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. doi:https://doi.org/10.1016/S1138-4891(12)70037-7

Tseng, F. M., & Hu, Y. C. (2010). Comparing four bankruptcy prediction models: logit, quadratic interval logit, neural and fuzzy neural networks. *Expert Systems with Applications*(37(3)), 1846--1853.

Zhang, J., Li, H., & Guo, Q. (2019). Bankruptcy prediction using deep neural networks: A survey and future research directions. *Expert Sustems with Applications*(115), 419-436.

ANEXO: Script de Python comentado

Enlace al código: Prediccion_quiebras_espanol_v1.py