Predicción de quiebras empresariales en turismo colombiano mediante k-NN funcional y una métrica de distancia personalizada (1995–2023)

Luis Eduardo Ruiz Paredes

Universidad EAN

Facultad de Administración, Finanzas y Ciencias Económicas

Doctorado en Gestión

Bogotá D.C., Colombia
2025

Predicción de quiebras empresariales en turismo colombiano mediante k-NN funcional y una métrica de distancia personalizada (1995–2023)

A Custom Functional Distance Metric and k-NN Classifier for Business Bankruptcy Prediction in Colombia's Tourism Sector (1995–2023)

Luis Eduardo Ruiz Paredes

Tesis doctoral presentada como requisito para optar al título de:

Doctor en Gestión

Director:

PhD. Hernando Porras Gómez

Universidad EAN

Facultad de Administración, Finanzas y Ciencias Económicas

Doctorado en Gestión

Bogotá D.C., Colombia
2025

Índice general

Índice de tablas

Índice de figuras

Resumen

Esta investigación doctoral desarrolla como aporte metodológico original una métrica funcional multivariada personalizada para comparar trayectorias financieras multivariadas de empresas. Esta métrica permite evaluar la similitud entre entidades a partir de la evolución de sus indicadores financieros durante cinco años, incluso en presencia de datos faltantes o escalas heterogéneas. Al integrarse en un clasificador k-vecinos más cercanos (k-NN), el modelo resultante ofrece un equilibrio entre rendimiento predictivo e interpretabilidad, permitiendo visualizar y analizar los vecinos más cercanos en el espacio funcional.

Para su validación empírica, se aplicó la metodología al problema de predicción de quiebra en empresas del sector turismo en Colombia, un segmento estratégico para la economía nacional que ha mostrado alta vulnerabilidad ante crisis financieras y operativas. Se construyó una base longitudinal con más de 5.000 empresas (1995–2023) y se identificaron eventos de quiebra mediante criterios financieros y operativos. A pesar del auge reciente de técnicas de aprendizaje automático aplicadas a este problema, persisten limitaciones clave: muchos modelos avanzados operan como "cajas negras" difíciles de interpretar, no permiten visualizar la similitud concreta entre empresas y tienden a ignorar la estructura dinámica de las trayectorias financieras multivariadas.

El modelo funcional fue evaluado con validación cruzada estratificada, alcanzando un F1-score promedio de 0,9802, una sensibilidad del 96,35 % y una precisión del 99,76 %, resultados que lo posicionan como una alternativa competitiva frente a modelos más complejos como XGBoost, redes LSTM o arquitecturas híbridas. Si bien no supera a todos en desempeño, su principal fortaleza radica en la posibilidad de identificar empresas similares (vecinas) de manera explícita y explicar la relevancia de cada indicador mediante los pesos optimizados en la métrica, lo que lo convierte en un complemento valioso

para sistemas de alerta temprana y análisis financiero.

Además del desarrollo metodológico, se construyó un aplicativo computacional interactivo que permite estimar el riesgo de quiebra de una empresa y mostrar sus vecinos financieros más cercanos en función de la métrica propuesta. Esta herramienta refuerza la aplicabilidad del modelo al proporcionar trazabilidad, explicaciones claras y posibilidades reales de uso por parte de analistas, supervisores o gestores no técnicos. La hipótesis de este trabajo plantea que un modelo k-NN funcional basado en una métrica de distancia personalizada, diseñada para trayectorias financieras multivariadas, puede alcanzar un desempeño competitivo frente a modelos avanzados, manteniendo interpretabilidad y capacidad de visualización de similitudes.

Palabras clave: Quiebras empresariales, predicción de quiebras, sector turismo, análisis funcional de datos, trayectorias multivariadas, métrica de distancia personalizada, clasificación funcional k-NN, modelos predictivos de riesgo.

Abstract

This doctoral research develops, as an original methodological contribution, a customized multivariate functional metric to compare multivariate financial trajectories of companies. This metric makes it possible to evaluate the similarity between entities based on the evolution of their financial indicators over five years, even in the presence of missing data or heterogeneous scales. When integrated into a k-nearest neighbors (k-NN) classifier, the resulting model offers a balance between predictive performance and interpretability, enabling the visualization and analysis of the closest neighbors in the functional space.

For its empirical validation, the methodology was applied to the problem of bankruptcy prediction in companies in the tourism sector in Colombia, a strategic segment for the national economy that has shown high vulnerability to financial and operational crises. A longitudinal database with more than 5,000 companies (1995–2023) was built, and bankruptcy events were identified using financial and operational criteria. Despite the recent rise of machine learning techniques applied to this problem, key limitations persist: many advanced models operate as "black boxes" that are difficult to interpret, do not allow for explicit visualization of similarity between companies, and tend to overlook the dynamic structure of multivariate financial trajectories.

The functional model was evaluated using stratified cross-validation, achieving an average F1-score of 0.9802, a sensitivity of 96.35 %, and a precision of 99.76 %. These results position it as a competitive alternative to more complex models such as XGBoost, LSTM networks, or hybrid architectures. Although it does not outperform all in performance, its main strength lies in the ability to explicitly identify similar (neighboring) companies and explain the relevance of each indicator through the optimized weights in the metric, making it a valuable complement for early warning systems and financial analysis.

In addition to the methodological development, an interactive computational application was built to estimate the bankruptcy risk of a company and display its closest financial neighbors based on the proposed metric. This tool reinforces the model's applicability by providing traceability, clear explanations, and real possibilities of use by analysts, supervisors, or non-technical managers. The hypothesis of this work states that a functional k-NN model based on a customized distance metric, designed for multivariate financial trajectories, can achieve competitive performance compared to advanced models, while maintaining interpretability and the ability to visualize similarities.

Keywords: Business bankruptcy, bankruptcy prediction, tourism sector, functional data analysis, multivariate trajectories, personalized distance metric, functional k-NN classification, predictive risk models.

Capítulo 1

Contextualización

1.1. Introducción

El enfoque propuesto en esta tesis es un clasificador k-NN funcional con métrica multivariada personalizada. Este método es valioso porque trabaja directamente con las trayectorias temporales completas de las variables financieras, capturando dinámicas de alta frecuencia que los modelos convencionales no explotan. En particular, Almanjahie et al. (2024) destacan que el k-NN funcional adapta localmente el ancho de banda a las curvas de datos (series financieras continuas), lo que mejora la identificación sistemática del riesgo financiero(Almanjahie et al., 2024). Asimismo, la métrica personalizada pondera distintas dimensiones de la trayectoria contable (por ejemplo, ratios financieros evolutivos) y, en su versión extendida, incorpora atributos categóricos relevantes como el sector o la localización, reforzando la capacidad del modelo para detectar señales relevantes. Bajo este marco, la investigación parte del supuesto de que una métrica funcional multivariada, capaz de penalizar las discrepancias estructurales entre trayectorias financieras, puede aumentar de manera significativa la precisión del modelo k-NN funcional sin perder su interpretabilidad. Este supuesto constituye la base de la pregunta y la hipótesis que orientan el desarrollo del trabajo, en torno a si es posible lograr un equilibrio real entre desempeño predictivo y trazabilidad metodológica en la predicción de quiebras empresariales. A diferencia de métodos "caja negra", el k-NN es inherentemente interpretable y trazable: cada predicción se fundamenta en ejemplos históricos reales, facilitando la comprensión y validación de los resultados. Cabe precisar que esta distancia

se plantea desde el inicio como una semi-métrica funcional penalizada, en coherencia con enfoques recientes para comparar series temporales complejas incluso con datos incompletos o choques estructurales (N. James et al., 2023).

La quiebra empresarial es un problema de gran relevancia tanto académica como práctica, potenciado por las recientes crisis económicas. En particular, el sector turismo –estratégico para la economía global y nacional– exhibe una alta vulnerabilidad financiera: estudios recientes señalan que la mayoría de las empresas turísticas analizadas operaban en situación crítica (por ejemplo, 96.4 % en la "zona de crisis" de Altman)(Nagendrakumar et al., 2023). En Colombia, la actividad turística alcanzó cifras históricas en 2023 (5.86 millones de visitantes internacionales y USD 8.547 millones en ingresos)(Téllez et al., 2024), lo que subraya la importancia de anticipar quiebras en este campo. Ante estos datos, el desarrollo de herramientas de alerta temprana para prevenir insolvencias resulta especialmente crítico.

La predicción de quiebra ha evolucionado metodológicamente en las últimas décadas. En sus inicios se basaba en modelos estadísticos clásicos (análisis discriminante, regresión logística/probit, Altman Z-score), pero en años recientes se han incorporado técnicas avanzadas de aprendizaje automático. Por ejemplo, Shetty et al. (2022) aplicaron XGBoost, SVM y redes neuronales profundas para predecir quiebras en PYMEs, alcanzando cerca de un 82–83 % de acierto con solo tres ratios financieros clave. De manera análoga, Pellegrino et al. (2024) propusieron una arquitectura LSTM multi-entrada que modela cada variable financiera por separado y demostraron que supera en precisión a enfoques tradicionales como la regresión logística o SVM. Además, se han explorado enfoques de supervivencia para incorporar la dimensión temporal del riesgo: Borges y Carvalho (2025) compararon modelos de riesgos proporcionales de Cox y Bosques Aleatorios de Supervivencia (Random Survival Forests) aplicados a pymes, encontrando que variantes avanzadas de RSF ofrecen mejor desempeño predictivo que el modelo de Cox clásico.

Esta tesis realiza explícitamente comparaciones entre el modelo funcional propuesto y otros métodos de referencia. Se evaluará su desempeño frente a modelos tradicionales (logit, probit, etc.) y avanzados (XGBoost, LSTM, RSF, entre otros), usando criterios de desempeño predictivo uniformes. De este modo se evidenciarán las ventajas e inconvenientes relativos de cada enfoque en el contexto de predicción de quiebras en el turismo colombiano.

En cuanto a la organización del documento, tras esta introducción, el **Capítulo 1** presenta la contextualización del problema, los objetivos, la hipótesis y la delimitación de la investigación. El **Capítulo 2** ofrece una

revisión exhaustiva del estado del arte sobre predicción de quiebra empresarial, con énfasis en modelos clásicos, avanzados y enfoques funcionales. El **Capítulo 3** desarrolla el marco metodológico del estudio, que constituye el núcleo conceptual y técnico de la tesis. En este capítulo se detalla la construcción del espacio funcional multivariado, el diseño de la métrica de distancia personalizada para trayectorias financieras, y su extensión para incorporar variables categóricas y temporales. A partir de este diseño metodológico, los siguientes capítulos se enfocan en su validación empírica. En particular, el **Capítulo 4** describe el proceso de consolidación, imputación y estructuración de la base de datos financiera utilizada, incluyendo el cálculo de indicadores y la construcción de la variable objetivo. El **Capítulo 5** presenta la evaluación empírica del modelo funcional, tanto en su versión base como en la versión extendida, con análisis de optimización, robustez y aplicación práctica. El **Capítulo 6** compara el modelo funcional con otros enfoques predictivos, incluyendo modelos tradicionales, avanzados, secuenciales e híbridos, bajo condiciones de comparación homogéneas. Finalmente, el **Capítulo 7** expone las conclusiones del estudio, los aportes al conocimiento y la práctica, así como las principales limitaciones y líneas de investigación futura.

1.2. Planteamiento del problema

En los últimos años, la predicción de quiebras empresariales ha estado marcada por un dilema metodológico: los modelos más avanzados en *machine learning* —como XGBoost, redes LSTM o enfoques híbridos— logran niveles de precisión muy altos, pero suelen operar como "cajas negras" con baja interpretabilidad. En contraste, métodos más simples como el k-Nearest Neighbors (k-NN) ofrecen un razonamiento más transparente, pues permiten explicar las predicciones mostrando casos históricos similares; sin embargo, su desempeño predictivo suele ser inferior al de esos modelos complejos. Este trabajo parte de ese contraste inicial y propone una solución que busca reducir la brecha: potenciar la capacidad predictiva de un k-NN funcional mediante una métrica de distancia especialmente diseñada para comparar trayectorias financieras, manteniendo —e incluso reforzando— su interpretabilidad (Gnip et al., 2025), (Guenani et al., 2024), (Dasilas & Rigani, 2024), (Qu et al., 2019), (Guenani et al., 2023).

La predicción temprana de la quiebra empresarial es un problema crítico en las finanzas, pues permite mitigar pérdidas y anticipar medidas correctivas para salvaguardar la continuidad de las empresas. En el caso del sector turismo en Colombia, compuesto mayormente por micro, pequeñas y medianas empresas (Romero Espinosa et al., 2015), esta cuestión cobra especial relevancia. Eventos recientes como la pandemia de COVID-19 han exacerbado la fragilidad financiera del sector: en 2020 se registró el mayor nivel de pérdidas de los últimos tres años, con 441 empresas turísticas acumulando pérdidas por COP 1,09 billones (Superintendencia de Sociedades, 2023). Esta brusca contracción de ingresos obligó a muchas compañías a buscar financiación para cubrir costos fijos, incrementando sus deudas (Romero Espinosa et al., 2015). Como resultado, el riesgo de quiebra empresarial en turismo se ha elevado significativamente, poniendo de manifiesto la necesidad de herramientas robustas que permitan predecir a tiempo cuáles empresas están en mayor peligro de insolvencia. Además, estudios recientes sobre sostenibilidad y percepción en destinos turísticos del Caribe colombiano subrayan la vulnerabilidad del sector y la urgencia de planificación estratégica e informada (Santiago et al., 2024).

Desde la perspectiva académica, la predicción de quiebra ha sido ampliamente estudiada durante décadas. Iniciativas clásicas como el modelo Z-score de Altman demostraron que es posible anticipar la quiebra mediante indicadores financieros combinados (Zhao et al., 2024b). A lo largo del

tiempo, los enfoques evolucionaron desde métodos estadísticos tradicionales (análisis discriminante múltiple, regresión logística, etc.) hacia técnicas de aprendizaje automático. La literatura reciente muestra una proliferación de modelos avanzados, incluyendo métodos de *ensemble* (por ejemplo, bosques aleatorios y XGBoost) y redes neuronales profundas (como las redes LSTM para secuencias), e incluso esquemas híbridos que combinan múltiples técnicas (Dasilas & Rigani, 2024; Guerra & Castelli, 2021). Estos modelos de última generación suelen lograr alta precisión predictiva en la detección de empresas en quiebra (Zhao et al., 2024b). Sin embargo, también presentan desafíos notorios. En primer lugar, a menudo requieren grandes volúmenes de datos históricos y variedad de variables para entrenar adecuadamente, lo cual puede ser un obstáculo en sectores con datos limitados o confidenciales (como muchas PYMEs de turismo). En segundo lugar —y más crucial aún—, tienden a comportarse como "cajas negras", con escasa interpretabilidad de sus resultados. De hecho, se ha observado que arquitecturas complejas como las redes recurrentes LSTM no facilitan la identificación de la importancia de cada variable explicativa en la predicción de la quiebra (Dasilas & Rigani, 2024), lo que dificulta que los analistas entiendan por qué un modelo está etiquetando a una empresa como riesgosa. Este déficit de transparencia genera preocupación entre los *stakeholders*, quienes requieren explicaciones claras para confiar en las alertas de riesgo y tomar decisiones informadas (Guerra & Castelli, 2021). La investigación reciente reconoce este problema de interpretabilidad; se han explorado técnicas como LIME (Local Interpretable Model-Agnostic Explanations) para intentar explicar las predicciones de modelos de quiebra basados en *machine learning* (Dasilas & Rigani, 2024). Pese a tales esfuerzos, la realidad es que los métodos más precisos suelen sacrificar claridad, creando una brecha entre desempeño predictivo e interpretabilidad que aún no ha sido resuelta del todo en el ámbito de la predicción de insolvencia empresarial.

Otro aspecto importante del problema es la naturaleza temporal del riesgo de quiebra empresarial. Muchos estudios previos simplifican el análisis considerando solo instantáneas estáticas (por ejemplo, ratios financieros de un único año) para predecir la quiebra. Sin embargo, las empresas atraviesan trayectorias financieras a lo largo del tiempo; es decir, sus indicadores contables y financieros evolucionan año tras año, mostrando tendencias, ciclos o señales de deterioro progresivo (Alaka et al., 2018; Shi & Li, 2019). Una serie de cinco años de datos financieros brinda información más rica que un solo año aislado, pues permite captar patrones dinámicos (como un aumento sostenido del endeudamiento, disminución paulatina de la liquidez, etc.) que suelen

preceder a una quiebra. Los modelos secuenciales modernos, como las redes LSTM, intentan precisamente explotar estas secuencias temporales, pero al costo ya señalado de introducir complejidad y opacidad en la interpretación (Qu et al., 2019). En el contexto colombiano y del sector turismo, aprovechar las trayectorias históricas resulta especialmente valioso debido a posibles particularidades en los ciclos económicos locales o estacionales propios del turismo, que un modelo estático podría pasar por alto.

En síntesis, el problema que se plantea es el siguiente: ¿De qué manera la incorporación de una métrica funcional personalizada, capaz de comparar trayectorias financieras multivariadas, puede mejorar el poder predictivo de un modelo k-NN funcional en la estimación del riesgo de quiebra de empresas del sector turismo en Colombia, hasta convertirlo en una alternativa competitiva frente a modelos avanzados (como XGBoost, redes LSTM y enfoques híbridos) sin sacrificar la interpretabilidad de los resultados? Existe la necesidad de un enfoque que logre un equilibrio entre un alto poder predictivo —comparable al de los métodos más avanzados reportados en la literatura— y la interpretabilidad y sencillez necesarias para su adopción práctica. Investigaciones recientes como las de Correa (2023) han mostrado que, en el contexto colombiano, modelos basados en técnicas como XGBoost pueden alcanzar altos niveles de precisión en la predicción de quiebra, pero presentan limitaciones en cuanto a explicabilidad. De forma similar, Yousaf et al. (2022) comparan modelos estáticos, dinámicos y de aprendizaje automático para empresas chinas, concluyendo que los enfoques basados en árboles como Random Forest y XGBoost superan en desempeño a los modelos tradicionales, aunque a costa de menor transparencia.

Frente a este dilema, surgen propuestas metodológicas que buscan potenciar modelos clásicos. Por ejemplo, X. Hu et al. (2022) proponen el uso del método k-Nearest Neighbors (k-NN) en un marco funcional no paramétrico para datos dependientes, destacando su flexibilidad y capacidad de adaptación a estructuras temporales complejas (X. Hu et al., 2022). De forma complementaria, Guenani et al. (2023), han demostrado que los estimadores kNN pueden robustecerse y adaptarse al contexto de datos funcionales ergódicos, con propiedades asintóticas bien establecidas (Guenani et al., 2023, 2024).

Esta investigación se inserta en ese marco y propone una solución innovadora pero conceptualmente simple: una métrica funcional personalizada que permita comparar objetivamente las trayectorias financieras multivariadas de diferentes empresas. Al integrarse esta métrica en un modelo clásico como el k-NN funcional, se busca capitalizar la información temporal de cinco

años de datos financieros por empresa y, al mismo tiempo, generar un modelo transparente cuyas predicciones puedan explicarse mostrando empresas históricamente similares (vecinos). De este modo, se pretende transformar un modelo básico en una alternativa competitiva frente a los enfoques más complejos, cerrando así la brecha entre interpretabilidad y precisión en la predicción de quiebras para el sector turismo en Colombia. En este punto surge la cuestión central que orienta la investigación, relacionada con la posibilidad de que una métrica funcional multivariada personalizada logre equilibrar precisión e interpretabilidad hasta volver competitivo al modelo k-NN frente a los métodos de vanguardia.

El desarrollo de esta tesis se estructura a partir de un conjunto de objetivos específicos que orientan tanto la construcción metodológica como la validación empírica del modelo funcional propuesto. Dichos objetivos articulan la formulación conceptual de la métrica personalizada con su posterior evaluación y comparación frente a otros modelos de predicción de quiebra.

1.3. Justificación

La justificación de este trabajo se sustenta en argumentos tanto teóricos como prácticos, que evidencian su relevancia académica y utilidad aplicada.

Contribución metodológica y brecha en la literatura

Académicamente, el estudio responde a una necesidad ampliamente reconocida en la literatura reciente: mejorar la predicción de quiebras integrando información temporal multianual sin incurrir en la opacidad de los modelos más complejos de *machine learning*. Diversos trabajos han demostrado que técnicas avanzadas como XGBoost, CatBoost y redes neuronales profundas logran altos niveles de precisión en la clasificación de empresas en riesgo (Ben Jabeur et al., 2021; Iparraguirre et al., 2024; Y. Liu et al., 2025). Sin embargo, también se ha señalado que estos métodos tienden a comportarse como "cajas negras", dificultando la comprensión del razonamiento detrás de las predicciones. P. Carmona et al. (2022), por ejemplo, argumentan que incluso modelos tan efectivos como XGBoost presentan barreras importantes para su interpretación por parte de usuarios no técnicos.

La métrica funcional personalizada desarrollada en esta tesis se plantea como una innovación metodológica dirigida específicamente a ese punto crítico: en lugar de proponer un algoritmo completamente nuevo, se refuerza un modelo clásico (el *k-Nearest Neighbors*) mediante la incorporación de una distancia definida sobre trayectorias financieras multivariadas. Este enfoque se inserta en el paradigma del Análisis de Datos Funcionales (FDA), extendiendo sus aplicaciones hacia la predicción de riesgo financiero con trayectorias reales.

De esta forma, se propone un modelo conceptualmente simple pero empíricamente robusto, que puede competir en rendimiento con los modelos más sofisticados, manteniendo una estructura comprensible para analistas e instituciones. Esta combinación de desempeño e interpretabilidad responde a una laguna persistente en la literatura, donde la elección entre precisión e inteligibilidad ha sido históricamente excluyente (Ben Jabeur et al., 2021). Además, el trabajo plantea una comparación sistemática con múltiples enfoques alternativos —desde regresión logística hasta arquitecturas híbridas basadas en texto y contabilidad— lo cual refuerza la solidez del aporte metodológico al mostrar su lugar relativo dentro del estado del arte.

Interpretabilidad y adopción tecnológica

Un elemento destacado de la justificación es la interpretabilidad mejorada que ofrece el enfoque propuesto. La posibilidad de explicar una predicción de quiebra mostrando las empresas vecinas más similares (aquellas cuya trayectoria financiera histórica se asemeja a la de la empresa analizada) constituye una ventaja crucial en entornos de decisión. Esta capacidad explicativa, propia del modelo funcional *k-NN*, ha sido destacada por Guenani et al. (2024), quienes demuestran que el método permite una estimación robusta y localmente adaptativa aún en contextos ergódicos, facilitando la trazabilidad de las predicciones.

Este tipo de explicabilidad fortalece la toma de decisiones informada: por ejemplo, si el modelo señala que la Empresa X presenta alto riesgo debido a similitudes con Empresas Y y Z —que efectivamente quebraron—, los directivos de X pueden identificar patrones compartidos (como un sobreendeudamiento creciente o la caída progresiva de los márgenes) y adoptar medidas correctivas con base en evidencia concreta. Además, estudios como el de X. Hu et al. (2022) han confirmado que el estimador funcional k-NN permite preservar la interpretabilidad incluso cuando se amplía a contextos dependientes mediante condiciones como asociación negativa.

La literatura enfatiza que un modelo de predicción de fracaso empresarial solo genera valor real cuando sus resultados pueden traducirse en acciones concretas de mejora o prevención. En este sentido, el uso de modelos funcionales explicables reduce la brecha entre analistas técnicos y tomadores de decisiones, promoviendo la adopción tecnológica (Guenani et al., 2023). Frente a la "caja negra" de modelos complejos como las redes neuronales, la transparencia del enfoque funcional con vecinos más cercanos puede mejorar significativamente la confianza y aceptación del modelo, especialmente entre usuarios no especialistas como gerentes o analistas de riesgo.

Desarrollo de un aplicativo y replicabilidad

Como parte de los aportes aplicados, la tesis incluye el desarrollo de un aplicativo computacional interactivo. Este software permite estimar el riesgo de quiebra de una empresa a partir de sus trayectorias financieras históricas y visualizar gráficamente los casos más similares, lo que refuerza la aplicabilidad práctica del modelo. Según P. Carmona et al. (2022), uno de los mayores desafíos de los modelos de aprendizaje automático es que, pese a su precisión,

suelen operar como "cajas negras" difíciles de interpretar para los usuarios finales. Al encapsular el modelo funcional *k-NN* en un aplicativo reutilizable, la investigación propone una solución accesible, transparente y transferible a distintos sectores.

Esta estrategia contribuye en dos frentes: primero, al facilitar la adopción práctica de herramientas predictivas avanzadas por parte de analistas o gestores sin formación técnica, y segundo, al permitir la extensibilidad del enfoque metodológico hacia otros dominios, como lo han planteado estudios recientes sobre replicabilidad en predicción de quiebras (Papík & Papíková, 2024a). Como enfatizan Papík y Papíková (2024a), el uso de modelos automatizados y reutilizables representa un avance significativo para extender el acceso a herramientas predictivas, reduciendo los costos de implementación y facilitando la adopción institucional.

En este sentido, el modelo propuesto se alinea con las tendencias actuales de ciencia de datos financiera, que valoran no solo la precisión y robustez del algoritmo, sino también su capacidad para ser compartido, auditado y replicado por terceros (Barboza & Altman, 2024). Al estar basado en un algoritmo clásico (*k-NN*) potenciado con una métrica funcional especializada, el enfoque garantiza replicabilidad incluso en entornos con recursos computacionales limitados, a diferencia de los modelos neuronales profundos cuyo reentrenamiento puede requerir condiciones técnicas difíciles de replicar.

En suma, esta combinación de métrica innovadora, modelo interpretable y herramienta funcional asegura un impacto real tanto en la academia como en la industria, alineándose con las exigencias de reproducibilidad y aplicabilidad propias de los desarrollos en predicción de quiebra empresarial.

1.4. Pregunta de investigación

A partir del problema descrito, se formula la siguiente pregunta de investigación:

¿En qué medida una métrica funcional multivariada personalizada puede potenciar el desempeño predictivo e interpretabilidad de un clasificador k-NN funcional frente a métricas y modelos avanzados existentes, validando su eficacia a través de un caso de estudio en el sector turismo colombiano?

Esta pregunta orienta el estudio hacia la búsqueda de un método que conjugue desempeño e interpretabilidad, evaluando concretamente si un enfoque basado en datos funcionales (trayectorias temporales) y aprendizaje estadístico clásico puede rivalizar con las técnicas de vanguardia comúnmente empleadas en la predicción de insolvencia empresarial.

1.5. Objetivos

Objetivo general

Desarrollar un enfoque metodológico para la predicción del riesgo de quiebra en empresas del sector turismo colombiano, basado en el análisis de trayectorias financieras multivariadas y en la definición de una métrica funcional personalizada integrada a un modelo k-NN clásico con alta capacidad predictiva e interpretativa.

Objetivos específicos:

1. Diseñar una métrica funcional multivariada personalizada que cuantifique la similitud entre las trayectorias financieras de diferentes empresas, integrando sus indicadores contables para detectar patrones relevantes de deterioro (tendencias, volatilidades y rupturas estructurales).

2. Implementar la métrica personalizada propuesta dentro de un clasificador k-NN funcional, adaptado al análisis de datos financieros temporales, de manera que el modelo pueda aprovechar dicha métrica para clasificar empresas según su riesgo de quiebra.

3. Preparar el conjunto de datos empíricos a partir de la información financiera reportada por la Superintendencia de Sociedades para empresas del sector turismo colombiano, mediante la estructuración de las trayectorias financieras y la identificación de eventos de quiebra con base en evidencia operativa y contable, garantizando la solidez metodológica del insumo utilizado en la validación del modelo.

4. Evaluar el desempeño predictivo del modelo k-NN funcional con la métrica funcional personalizada, mediante técnicas de validación cruzada, métricas de evaluación robustas (F1-score, AUC y precisión) y análisis de robustez estadística, asegurando la consistencia y fiabilidad de los resultados obtenidos en la clasificación de empresas según su riesgo de quiebra.

5. Comparar los resultados predictivos del modelo propuesto con los de enfoques de referencia, incluyendo modelos estadísticos tradicionales (regresión logística), métodos de aprendizaje automático modernos

(XGBoost) y redes neuronales secuenciales (LSTM), con base en la evaluación de diferencias estadísticas en precisión e interpretabilidad.

6. Desarrollar un aplicativo computacional para la estimación del riesgo de quiebra de empresas del sector turismo colombiano, basado en la métrica funcional personalizada y el modelo k-NN funcional, con integración de componentes visuales que muestren los casos más similares y faciliten la interpretación de las predicciones por parte del analista.

1.6. Hipótesis

En consonancia con la pregunta de investigación, se plantea la siguiente hipótesis:

Una métrica funcional personalizada, capaz de capturar la similitud entre trayectorias financieras multivariadas, otorgará a un modelo k-NN funcional un poder predictivo significativamente superior al de su contraparte con métricas convencionales, hasta alcanzar un desempeño equiparable al de modelos de alta complejidad (como XGBoost, LSTM y modelos híbridos) en la predicción de la quiebra de empresas del sector turismo colombiano.

Asimismo, se espera que este modelo conserve una mayor interpretabilidad, ya que sus predicciones podrán explicarse mediante la comparación directa con empresas vecinas que presenten características financieras similares.

En términos específicos, la hipótesis implica que no existirá una diferencia estadísticamente significativa en las métricas de desempeño (por ejemplo, F1-score, AUC, precisión) entre el k-NN funcional potenciado con la nueva métrica y los modelos avanzados de referencia. A su vez, el modelo k-NN funcional proporcionará puntos interpretativos más claros, como la identificación de patrones financieros análogos en empresas vecinas que sirvan como evidencia para justificar por qué se predice la quiebra o la supervivencia.

Esta hipótesis combina dos dimensiones —eficacia predictiva e interpretabilidad— y orienta la validación empírica del enfoque propuesto.

1.7. Delimitación de la investigación

Esta investigación se encuentra delimitada en varios aspectos fundamentales que acotan su alcance, tanto en términos metodológicos como empíricos, con el fin de garantizar coherencia interna y claridad en la interpretación de los resultados.

- **Delimitación temática:** El estudio se enfoca exclusivamente en la predicción del riesgo de quiebra empresarial a partir del análisis de trayectorias financieras multivariadas, dejando de lado aspectos causales, normativos o cualitativos asociados a la quiebra. El objetivo es evaluar si la incorporación de una métrica funcional personalizada mejora la capacidad predictiva y la interpretabilidad de un modelo k-NN funcional en comparación con enfoques tradicionales y avanzados.

- **Delimitación geográfica:** La población de estudio está compuesta únicamente por empresas del sector turismo que operan en Colombia y que reportan información financiera a la Superintendencia de Sociedades. No se incluyen empresas informales ni sectores económicos distintos.

- **Delimitación temporal:** Si bien la base de datos incluye información financiera entre 1995 y 2023, la construcción de las trayectorias no exige la disponibilidad ininterrumpida de cinco años consecutivos, permitiendo la inclusión de empresas con trayectorias parcialmente incompletas, siempre que cumplan con criterios mínimos de información válida.

- **Delimitación poblacional:** La muestra está compuesta por empresas formales registradas ante la Superintendencia de Sociedades y clasificadas dentro del sector turismo, de acuerdo con el código CIIU. La variable dependiente utilizada es una medida binaria de riesgo de quiebra, construida a partir de criterios financieros objetivos.

- **Delimitación de variables:** Inicialmente se consideraron 45 indicadores financieros, pero mediante un proceso de selección basado en análisis de agrupamiento jerárquico y el factor de inflación de la varianza (VIF), se redujo el conjunto a 17 variables representativas. No se incorporaron variables categóricas, institucionales o macroeconómicas, ya que el propósito del estudio fue evaluar exclusivamente si las trayectorias financieras por sí solas pueden ofrecer suficiente poder predictivo.

- **Delimitación metodológica:** El modelo principal es un clasificador k-NN funcional alimentado por trayectorias financieras y una métrica personalizada que penaliza la pérdida de información temporal. Se compara su desempeño con modelos tradicionales (regresión logística, modelo Probit), avanzados (Random Forest, XGBoost) y secuenciales (LSTM), evaluando tanto el poder predictivo como la interpretabilidad. La investigación se circunscribe al análisis supervisado, sin aspiraciones causales.
- **Delimitación instrumental:** La implementación se realizó completamente en Python, empleando técnicas de validación cruzada y optimización de hiperparámetros mediante la biblioteca `Optuna`. Las bases de datos utilizadas fueron consolidadas, limpiadas y validadas previamente, y todas las transformaciones están documentadas para asegurar la trazabilidad del proceso.

Capítulo 2

Marco teórico y estado del arte en la predicción del riesgo empresarial

2.1. Fundamentos metodológicos de los modelos de aprendizaje automático para predicción de quiebra

El riesgo de quiebra se refiere a la probabilidad de que una empresa sea incapaz de cumplir sus obligaciones financieras, entrando en estado de insolvencia. En la literatura reciente se define como la posibilidad de que la empresa enfrente un procedimiento formal de bancarrota o incumpla sus deudas (Horváthová & Mokrišová, 2018). Este concepto abarca la incapacidad de pago de deudas (default) e implica consecuencias financieras adversas, por lo que su predicción temprana es crucial para evitar pérdidas mayores. Dada la importancia de anticipar la insolvencia para la toma de decisiones de inversores, gerentes y reguladores, múltiples estudios han explorado modelos que detectan señales de alerta temprana de problemas financieros (Issa et al., 2024).

Desde el punto de vista metodológico, el análisis de datos funcionales (FDA) provee un marco natural para modelar series temporales financieras como funciones continuas. Los datos funcionales interpretan cada observación

como una trayectoria suave en el tiempo (por ejemplo, una curva de ratios financieros a lo largo de varios años) (J.-L. Wang et al., 2016).

Formalmente, en el caso multivariado, cada empresa se representa mediante un vector de funciones

$$X = \left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right),$$

donde cada $X^{(j)} : T_j \to \mathbb{R}$ representa la trayectoria temporal de un indicador financiero (como flujos de caja, liquidez o rentabilidad), definida sobre un dominio temporal T_j y considerada como función medible.

El espacio funcional multivariado se define de forma general como el producto cartesiano de espacios funcionales individuales:

$$\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_p,$$

donde cada $\mathcal{F}_j$ es un subconjunto adecuado de funciones medibles sobre T_j, dependiendo del tipo de métrica utilizada. Por ejemplo, si se trabaja con la métrica euclidiana funcional inducida por la norma L^2, es necesario que cada $X^{(j)} \in L^2(T_j)$, es decir, que sea cuadrado-integrable:

$$\int_{T_j} \left|X^{(j)}(t)\right|^2 dt < \infty.$$

Esto da lugar al espacio de Hilbert multivariado:

$$\mathcal{H} = L^2(T_1) \times \cdots \times L^2(T_p),$$

dotado del producto interno:

$$\langle f, g \rangle = \sum_{j=1}^{p} \int_{T_j} f^{(j)}(t)\, g^{(j)}(t)\, dt, \quad \text{con norma} \quad \|f\| = \sqrt{\langle f, f \rangle}.$$

Sin embargo, en otros enfoques —como aquellos basados en distancias tipo L^1, métricas penalizadas, transformaciones acotadas o semimétricas construidas para datos ruidosos o incompletos— los requisitos pueden ser menos estrictos. En estos casos, puede bastar con que las trayectorias tengan una distancia finita respecto a otras funciones del espacio, sin requerir necesariamente cuadrado-integrabilidad (M. G. Chen, 2023).

Esta estructura permite capturar la evolución temporal completa de las variables financieras de la empresa como trayectorias continuas, aprovechando

la suavidad inherente de datos de alta frecuencia (p. ej., cotizaciones diarias, que son casi continuas) para una modelación eficaz (Horváthová & Mokrišová, 2018; J.-L. Wang et al., 2016). Así, en lugar de analizar valores aislados o ventanas de tiempo discretas, FDA trata las series de tiempo completas como objetos funcionales, facilitando técnicas como componentes principales funcionales o regresión funcional para estudiar patrones de comportamiento a lo largo del ciclo de vida de la empresa.

En este contexto, el clasificador k-NN funcional extiende el algoritmo clásico de k-vecinos más cercanos al espacio de funciones. Dado un conjunto de entrenamiento compuesto por observaciones funcionales etiquetadas (por ejemplo, empresas en quiebra o solventes) y una nueva trayectoria por clasificar, el método asigna la etiqueta mayoritaria entre los k vecinos más cercanos de la nueva muestra, medidos según una métrica adecuada definida en el espacio funcional $\mathcal{H}$ (Baíllo et al., 2011).

Es decir, se define una función de distancia $d(f, g)$ sobre funciones (por ejemplo, la norma L^2) y se seleccionan las k observaciones con menor distancia respecto a la muestra dada. En la práctica, se suele utilizar la distancia L^2 multivariada, definida como:

$$d_{L^2}(f,g) = \left(\sum_{j=1}^{p} \int_{T_j} \left(f^{(j)}(t) - g^{(j)}(t) \right)^2 dt \right)^{1/2}.$$

Luego, la regla de clasificación correspondiente es:

$$\hat{y}(x) = \text{mode}\left\{ y^{(1)}, y^{(2)}, \ldots, y^{(k)} \right\},$$

donde $y^{(i)}$ representa la clase de la i-ésima observación más cercana en distancia a x (Baíllo et al., 2011). Como en el k-NN clásico, este método es no paramétrico y perezoso (lazy): no requiere entrenamiento previo, sino que compara directamente la nueva trayectoria con las del conjunto de entrenamiento. Sin embargo, el desempeño del clasificador k-NN funcional depende fuertemente de la elección de la métrica funcional y de la calidad de los datos. Por ejemplo, la distancia L^2 estándar asume funciones definidas sobre el mismo dominio temporal y completamente observadas. Esta métrica, conocida como distancia Euclídea funcional, extiende la noción clásica al caso continuo mediante la expresión:

$$d_{L^2}(i,j) = \left(\int_T [x_i(t) - x_j(t)]^2 \, dt \right)^{1/2} \tag{2.1.1}$$

Aunque es conceptualmente directa, penaliza fuertemente las discrepancias puntuales, lo que puede amplificar el efecto de valores atípicos. Ha sido utilizada, por ejemplo, en contextos ergódicos por Guenani et al. (2023).

Una alternativa más robusta frente a *outliers* es la distancia Manhattan funcional, basada en la norma L^1, definida como:

$$d_{L^1}(i,j) = \int_T |x_i(t) - x_j(t)|\, dt \tag{2.1.2}$$

Este enfoque ha demostrado ser más tolerante a episodios extremos y ha sido aplicado en modelos funcionales robustos como los de Almanjahie et al. (2024).

Otra métrica común en análisis longitudinal es la basada en correlación o forma, que evalúa la similitud estructural entre trayectorias mediante $1 - \rho_{ij}$, siendo ρ_{ij} el coeficiente de correlación entre las funciones. Este enfoque enfatiza los patrones compartidos más que la magnitud absoluta, y ha sido empleado en entornos económicos por Park et al. (2021).

Por otro lado, el *Dynamic Time Warping* (DTW) permite alinear series temporales que presentan desfases, ajustando la comparación a diferencias en la velocidad de evolución. Aunque no constituye una métrica estricta en el sentido matemático, es ampliamente utilizada por su eficacia en contextos financieros, como lo evidencian estudios recientes como W. Chen et al. (2025).

Una opción más sofisticada es la semimétrica de Mahalanobis funcional, que incorpora la estructura de covarianza temporal entre trayectorias. Galeano et al. (2015) propusieron una versión que proyecta las funciones en bases ortogonales, logrando una discriminación más sensible al contexto dinámico de los datos.

Finalmente, existen otras métricas especializadas, como aquellas basadas en derivadas, que capturan la velocidad de cambio de los indicadores financieros; medidas de profundidad funcional (*depth measures*), que permiten comparar centralidad relativa de curvas; o distancias de tipo Fréchet, que consideran deformaciones espaciales y temporales de forma conjunta. Estas variantes son particularmente útiles para identificar patrones de deterioro acelerado o trayectorias atípicas previas a la quiebra.

Además, existen distancias elásticas penalizadas que incorporan un costo por dominios no observados o por lagunas en las trayectorias: estas métricas penalizan explícitamente las regiones con datos faltantes o caminos de alineación excesivamente largos, adaptándose mejor a series temporales con valores perdidos o asimetrías en el horizonte temporal.

En general, la métrica elegida debe reflejar una noción significativa de similitud entre trayectorias financieras, tolerando desplazamientos temporales o vacíos de datos, como ocurre en métodos de imputación funcional y en editores elásticos de curvas.

Cabe resaltar que el clasificador k-NN funcional puede ser computacionalmente costoso y sensible al ruido o al desbalance de clases (Amer et al., 2025), por lo que en la práctica se suele complementar con técnicas de reducción de dimensiones o selección de vecinos.

En el contexto de datos financieros funcionales, elegir la métrica adecuada es crucial. Las métricas L^2 y L^1 requieren trayectorias bien alineadas, mientras que DTW y la alineación por eventos críticos (como el año de quiebra) mejoran la comparabilidad en series desfasadas. Algunas investigaciones recientes han explorado combinaciones ad hoc: por ejemplo, ponderar más los años previos a la quiebra, como en Alamari et al. (2024).

En suma, si la métrica captura adecuadamente la dinámica financiera relevante, el modelo k-NN funcional puede superar las limitaciones del k-NN convencional y alcanzar niveles de desempeño comparables a modelos complejos como XGBoost o LSTM, sin perder interpretabilidad. Desarrollo de métricas funcionales para trayectorias financieras

Para dimensionar los modelos y comparar enfoques, la literatura considera también modelos avanzados modernos. Por ejemplo, *XGBoost* y *LightGBM* son algoritmos de *gradient boosting* basados en árboles de decisión: producen clasificadores altamente precisos mediante la combinación de múltiples árboles débiles T. Chen y Guestrin (2016) y Ke et al. (2017). Aunque su estructura permite calcular medidas de importancia de variables (por ejemplo, por ganancia o cobertura), estas no siempre reflejan correctamente el aporte real de cada predictor. Por ello, en aplicaciones sensibles como la predicción de quiebra, se suelen complementar con técnicas más robustas de interpretabilidad como *SHAP* o *LIME*, que permiten descomponer las predicciones en contribuciones individuales de cada variable Lundberg y Lee (2017) y Ribeiro et al. (2016).

Sin embargo, en su forma estándar estos modelos no capturan directamente la dependencia temporal, ya que requieren que las series se conviertan en características estáticas (por ejemplo, utilizando rezagos o agregaciones).

Por su parte, los modelos de *Long Short-Term Memory* (LSTM) son redes neuronales recurrentes diseñadas para procesar secuencias: incorporan celdas de memoria que retienen dependencias de largo plazo, lo que permite aprender patrones temporales complejos directamente a partir de las series financieras Hochreiter y Schmidhuber (1997). En contraste con los modelos

basados en árboles, las LSTM tienden a comportarse como cajas negras de difícil interpretación, aunque muestran un desempeño destacado en tareas puramente secuenciales.

En síntesis, la aproximación funcional *k-NN* destaca por su simplicidad y la capacidad de usar las trayectorias completas como objetos de comparación, mientras que modelos como *XGBoost*, *LightGBM* o *LSTM* exigen una parametrización más compleja. Los métodos de *boosting* tienden a ofrecer alto rendimiento predictivo al combinar múltiples árboles débiles sobre características financieras estáticas, y permiten cierta interpretabilidad mediante técnicas como *SHAP* (Correa, 2023). Sin embargo, no explotan explícitamente la naturaleza secuencial de las series temporales. Por su parte, los modelos *LSTM* modelan directamente la dimensión temporal y aprenden patrones dinámicos complejos, aunque suelen comportarse como cajas negras de difícil interpretación y requieren volúmenes significativos de datos.

El clasificador funcional *k-NN*, en cambio, proporciona una alternativa no paramétrica que combina la flexibilidad del análisis de datos funcionales con la interpretabilidad de los clasificadores basados en distancia. No obstante, su uso ha sido escaso en la literatura reciente sobre predicción de quiebra, en parte por su mayor costo computacional y por la necesidad de definir métricas específicas adaptadas a trayectorias financieras (Dasilas & Rigani, 2024; Gnip et al., 2025). Por tanto, su aplicación resulta más adecuada en contextos donde se privilegia la trazabilidad del modelo y se dispone de una métrica funcional bien fundamentada.

2.2. Estado del arte

La predicción de quiebra empresarial tiene sus raíces en modelos estadísticos clásicos. Un trabajo pionero de Altman (1968) aplicó el análisis discriminante múltiple sobre ratios financieros para anticipar insolvencias, sentando las bases de los modelos de scoring financiero. Posteriormente, Ohlson (1980) introdujo la regresión logística para la predicción de quiebras, incorporando probabilidades de fracaso empresarial en lugar de un simple score. Durante las décadas siguientes proliferaron métodos estadísticos tradicionales, pero a partir de los 2000 hubo un giro hacia técnicas de machine learning. Por ejemplo, redes neuronales artificiales, árboles de decisión, support vector machines (SVM) y métodos de ensamblado empezaron a complementar o reemplazar a los modelos lineales. Un análisis de 49 estudios (2010–2015) realizado por Alaka et al. (2018) resume ocho técnicas populares —dos métodos estadísticos (discriminante y logística) y seis de inteligencia artificial (redes neuronales, SVM, rough sets, razonamiento basado en casos, árboles de decisión y algoritmos genéticos)— evaluadas bajo 13 criterios de desempeño (precisión, transparencia, requerimiento de selección de variables, manejo de tamaños de datos, etc.). Sus hallazgos mostraron que ninguna herramienta domina todas las dimensiones de rendimiento y que la integración informada de múltiples técnicas (modelos híbridos) suele proporcionar resultados más sólidos. En la misma línea, estudios comparativos recientes confirman que los algoritmos de aprendizaje automático superan a las metodologías tradicionales basadas únicamente en ratios financieros simples.

Franco et al. (2022) realizó un análisis bibliométrico de 327 publicaciones y evidenció que enfoques modernos como XGBoost, SVM, métodos de remuestreo como SMOTE, random forest (RF) y árboles de decisión (DT) logran una capacidad predictiva muy superior a la de los modelos tradicionales, especialmente cuando se trata de anticipar la quiebra con suficiente antelación temporal. Esta superioridad de las técnicas de machine learning se atribuye tanto a su mayor precisión como a la posibilidad de incorporar variables financieras y no financieras, las cuales contribuyen significativamente a mejorar las predicciones. En efecto, la inclusión de datos cualitativos (p. ej., calidad de la gestión, reputación, gobierno corporativo) junto con ratios financieros ha ganado tracción en años recientes, permitiendo una evaluación más integral de la salud de la empresa (Dasilas & Rigani, 2024).

En cuanto al desempeño y las tendencias, las revisiones sistemáticas recientes reflejan la evolución del campo. Kuizinienė et al. (2022) analizan 232

estudios sobre predicción de distress financiero con técnicas de inteligencia artificial, abarcando desde la preparación de datos (desbalance de clases, reducción de dimensionalidad) hasta los algoritmos más usados y métricas de rendimiento. Concluyen que es crítico balancear los datos —dada la escasez relativa de quiebras— y aplicar técnicas de reducción de atributos para mejorar la eficacia de los modelos, señalando también direcciones futuras poco exploradas. Por su parte, Cheraghali y Molnár (2024) presentan una revisión sistemática centrada en metodologías empleadas para predecir la quiebra de PYMES, recopilando 145 estudios entre 1973 y 2023. Identifican más de 1200 factores considerados en la literatura y hasta 80 métodos de estimación distintos, incluyendo técnicas de remuestreo, selección de variables y validación cruzada. Esta diversidad de enfoques refleja un campo altamente experimental, donde se exploran combinaciones de variables y algoritmos según el contexto. Una constante reciente es la proliferación de modelos híbridos que combinan metodologías múltiples. Diversos autores han reportado que la hibridación —por ejemplo, redes neuronales con algoritmos evolutivos, o modelos estadísticos con boosting— tiende a superar a enfoques aislados. En un estudio comparativo, los enfoques híbridos alcanzaron una exactitud cercana al 94 %, superior a métodos de IA no combinados (88 %) y a métodos estadísticos clásicos (81 %) Cheraghali y Molnár (2024). Así, la tendencia actual apunta a aprovechar fortalezas complementarias de distintos modelos, a la vez que se abordan desafíos persistentes como el desbalance de clases mediante técnicas especializadas como SMOTE. También se observa una preocupación creciente por emplear múltiples métricas de evaluación más allá de la exactitud, como sensibilidad, especificidad, curva ROC (AUC) y otras medidas. En suma, las investigaciones desde 2020 destacan el uso de modelos más sofisticados (ensambles, híbridos, deep learning) y el uso de información más rica, confirmando un progreso sustancial respecto a los enfoques clásicos puramente estáticos (Dasilas & Rigani, 2024).

Limitaciones del k-NN tradicional

.

El algoritmo k-Nearest Neighbors (k-NN) depende críticamente de la métrica de distancia elegida para medir la similitud entre empresas. Habitualmente se usa la distancia Euclídea en un espacio de variables financieras estáticas (por ejemplo, ratios de un año específico), lo cual asume que todas las variables contribuyen por igual y de forma lineal a la noción de cercanía.

Esto puede ser problemático: si algunas características no son informativas o están escaladas de forma diferente, la distancia Euclídea estándar diluirá la influencia de rasgos realmente relevantes. A diferencia de métodos como la regresión (que puede asignar coeficientes) o los árboles (que seleccionan variables discriminativas), el k-NN no realiza ponderación explícita de atributos –cada dimensión aporta por su magnitud en la distancia–, volviéndolo sensible a ruido y variables irrelevantes.

Guenani et al. (2023) han señalado que esta debilidad hace que el modelo sea particularmente vulnerable en contextos de alta colinealidad o heterogeneidad de escalas. A su vez, W. Hu (2022) destacan que, aunque el k-NN puede extenderse a escenarios no paramétricos robustos, su uso típico en quiebra empresarial sigue limitado por su rigidez en la representación de similitud. Estudios comparativos han notado que, sin un preprocesamiento riguroso o elección adecuada de métricas, el k-NN tiende a quedar rezagado frente a técnicas más sofisticadas. Zapata y Mukhopadhyay (2024) comparan 16 clasificadores, desde logit hasta modelos en ensamblaje, y observan que el k-NN estándar no se destaca entre los mejores debido a sus limitaciones inherentes, principalmente por la falta de mecanismos de aprendizaje interno (como ajuste de pesos o selección de variables).

Otra restricción importante es la dimensionalidad. En problemas de predicción de quiebra, a menudo se consideran decenas de variables financieras simultáneamente; si además se incorporan múltiples periodos de tiempo como características separadas, el vector descriptor de cada empresa se expande significativamente. El fenómeno conocido como *curse of dimensionality* provoca que, en espacios de alta dimensión, prácticamente cualquier punto (empresa) esté lejos de todos los demás, dificultando que la noción de "vecino más cercano" sea significativa. Esto puede degradar la eficacia de k-NN cuando se usan muchos indicadores financieros a la vez o series de tiempo "apiladas" como atributos estáticos. Asimismo, los datos financieros suelen presentar distribuciones sesgadas y valores extremos (outliers); la distancia Euclídea penaliza fuertemente diferencias grandes en una sola variable, de modo que una empresa con un valor atípico en cierta ratio podría aparecer demasiado lejana de sus pares, incluso si en los demás indicadores es similar. El k-NN tradicional carece de mecanismos internos para lidiar con outliers o escalas heterogéneas (más allá de normalizar datos previamente). Además, no aprovecha ninguna suposición o estructura del problema: es un método no paramétrico totalmente dependiente de los datos de entrenamiento almacenados. Esto implica también un coste computacional elevado cuando el conjunto

de referencia es grande, pues para clasificar un nuevo caso se deben calcular distancias a todos (o la mayoría) de los casos históricos. En contextos de miles de empresas, este coste puede ser significativo, aunque es manejable con las capacidades computacionales actuales (Alamari et al., 2024; Almanjahie et al., 2024; Guenani et al., 2023, 2024).

Finalmente, enfocado a la predicción de quiebras, el modelo *k-Nearest Neighbors* (k-NN) convencional no modela explícitamente la secuencia temporal de los datos financieros. Si una empresa tiene datos de varios años, una estrategia simple consiste en tomar promedios o valores finales como variables estáticas; sin embargo, esto implica perder información sobre la trayectoria de la empresa (por ejemplo, si sus finanzas van empeorando de forma gradual o abrupta).

Una alternativa es incluir los valores de cada año como variables separadas (por ejemplo, $X1_{\text{año1}}, X1_{\text{año2}}, \ldots, Xn_{\text{año2}}$), pero en ese caso la distancia euclídea trata cada año de forma independiente y no captura patrones temporales. Además, esta estrategia incrementa considerablemente la dimensionalidad, como se señaló previamente.

En resumen, el k-NN "base" tiene dificultades para capturar dinámicas temporales y relaciones complejas entre variables, a menos que se extienda o complemente su manera de calcular distancias. Estas limitaciones explican por qué, si bien el razonamiento basado en casos fue incluido entre las técnicas prometedoras en revisiones como la de Alaka et al. (2018), la literatura reciente ha privilegiado otros algoritmos más adaptados a datos financieros multidimensionales y longitudinales.

Estudios más recientes como las de Alamari et al. (2024) y Almanjahie et al. (2024) han comenzado a explorar variantes funcionales del k-NN que permiten tratar series temporales completas como objetos matemáticos continuos, en lugar de vectores discretos. Este enfoque funcional posibilita definir distancias directamente sobre trayectorias financieras, capturando su forma, evolución y regularidad a lo largo del tiempo. Tal como señalan Guenani et al. (2024), este tipo de extensión no solo mejora la sensibilidad del clasificador a las dinámicas temporales, sino que también puede alcanzar propiedades de convergencia asintótica bajo ciertas condiciones ergódicas, ampliando así el campo de aplicación del k-NN en contextos económicos complejos.

Otro desarrollo importante es el de métricas que consideren explícitamente la no linealidad y posibles cambios de régimen en las finanzas de la empresa. Por ejemplo, se han propuesto medidas que identifican eventos financieros clave (como caída abrupta de ingresos o aumento de endeudamiento) y comparan la

secuencia de estos eventos más que los valores numéricos crudos. Este enfoque, aún emergente, puede conceptualizar la similitud financiera en términos de "patrones críticos compartidos" Alamari et al. (2024).

Un aspecto distintivo de los datos financieros funcionales es que muchas trayectorias se alinean respecto a un evento de quiebre ($t = 0$), lo que facilita comparaciones horizonte-a-horizonte (por ejemplo, de $t = -5$ a $t = 0$). Esto permite utilizar métricas como L^2 o L^1 directamente, pero también se ha sugerido introducir ponderaciones temporales que aumenten el peso de los años cercanos a la quiebra. Formalmente, una métrica ponderada se expresaría como:

$$d_{wL2}(i, j) = \sqrt{\int [x_i(t) - x_j(t)]^2 w(t)\, dt},$$

donde $w(t)$ es una función creciente en la cercanía a $t = 0$, que puede ser calibrada para maximizar la separabilidad entre empresas quebradas y no quebradas Guenani et al. (2023).

Desde el punto de vista computacional, calcular distancias funcionales para miles de empresas puede ser costoso. Por ello, se ha propuesto reducir dimensionalidad usando análisis de componentes principales funcionales (FPCA). Esto permite representar cada trayectoria mediante pocos coeficientes ortogonales y luego aplicar métricas simples (como Euclídea o Mahalanobis) en ese subespacio. Según Galeano et al. (2015), esta combinación mejora la discriminación del clasificador k-NN y reduce el ruido.

La utilidad práctica del enfoque funcional se ha evidenciado en estudios recientes de predicción de quiebra: por ejemplo, X. Chen et al. (2025) hallaron que un preprocesamiento funcional robusto mejora el rendimiento de clasificadores, mientras que Park et al. (2021) resaltaron cómo ciertas trayectorias revelan patrones identificables antes del colapso financiero. Todo esto sugiere que una métrica bien diseñada puede capturar la "firma dinámica" de una empresa en crisis.

Comparación de Modelos Predictivos de Quiebra Empresarial

.

Modelos Estadísticos Clásicos

Regresión Logística Estática y Dinámica La regresión logística ha sido uno de los métodos tradicionales más utilizados para predecir quiebras corporativas desde los años 1980, tras los primeros modelos de análisis discriminante en los 60 (p.ej., el Z-score de Altman en 1968, Altman (1968)). La regresión logística (y su variante probit) convirtió la predicción de quiebra en un problema de clasificación binaria, generando probabilidades (O-score de Ohlson, 1980). Sus fortalezas incluyen la simplicidad e interpretabilidad: los coeficientes estimados indican la contribución de cada ratio financiero al riesgo de quiebra, lo cual facilita la comprensión económica. Además, requiere muestras relativamente pequeñas y proporciona estimaciones probabilísticas bien calibradas cuando el modelo está especificado correctamente (Campbell et al. (2008)).

Sin embargo, presenta limitaciones importantes en datos financieros longitudinales: asume una relación lineal logit entre ratios (que puede no captar patrones complejos o interacciones no lineales) y típicamente se aplica a snapshots estáticos (p.ej., datos de un año) ignorando la dinámica temporal. Incluso cuando se incluyen ratios de varios años como variables, la logística convencional no modela explícitamente la dependencia temporal entre observaciones sucesivas de una empresa. Esto puede llevar a perder información secuencial relevante, especialmente en datos de series temporales financieras donde las tendencias o aceleraciones en indicadores anticipan la quiebra.

Para incorporar la dimensión temporal, surgieron modelos de regresión logística dinámica, fundamentalmente equivalentes a modelos de riesgo discreto (hazard models). Un referente es el modelo de Shumway (2001), que propone reestructurar los datos en formato panel (empresa-año) e incluir la historia temporal para estimar probabilidades de quiebra condicionadas a la supervivencia previa. Estudios posteriores como Campbell et al. (2008) muestran que este enfoque dinámico usando tanto variables contables como de mercado tiene un alto poder explicativo y logra probabilidades pronosticadas bien calibradas históricamente. En otras palabras, la logística dinámica aprovecha información longitudinal (ej. caídas sucesivas de liquidez o rentabilidad) para mejorar la detección temprana de quiebras, superando a las regresiones estáticas que usan solo el último año. Ventajas clave de la logística dinámica son su base teórica sólida en modelos de duración y su capacidad de actualizar el riesgo a medida que nuevas observaciones anuales aparecen, capturando efectos temporales persistentes (p. ej., las probabilidades de default tienden a ser persistentes en el tiempo si una empresa se deteriora). No obstante,

sus desventajas incluyen la necesidad de asumir independencia condicional entre observaciones anuales (lo cual puede violarse si hay autocorrelación no capturada) y sigue imponiendo una forma funcional global lineal en los parámetros. Además, tanto la logística estática como la dinámica suelen verse afectadas por el serio desbalance de clases típico de la quiebra (muy pocas empresas quebradas frente a muchas sanas), requiriendo técnicas de remuestreo o ajuste de umbrales para mejorar su sensibilidad.

Máquinas de Vectores de Soporte (SVM)

Las máquinas de vectores de soporte (SVM) irrumpieron en la literatura de quiebra a inicios de los 2000 como una técnica prometedora por su éxito en tareas de clasificación. Una SVM busca el hiperplano que maximiza el margen entre clases (quiebra vs. no quiebra) en un espacio de características, pudiendo emplear kernels para transformar no linealmente los datos de entrada. En predicción de quiebra, las SVM demostraron ventajas tempranas: Tserng et al. (2011) aplicaron una SVM a contratistas en el sector construcción y observaron que superaba a modelos tradicionales, gracias a su capacidad para generalizar mejor en muestras pequeñas. Este resultado fue reforzado por Erdogan (2012), quienes mostraron que las SVM tienen un buen equilibrio entre precisión y generalización, especialmente cuando el conjunto de entrenamiento es limitado. Esto se debe a que las SVM, al maximizar el margen, tienden a evitar el sobreajuste más eficazmente que redes neuronales con complejidad similar.

Además, las SVM pueden manejar eficientemente múltiples variables mediante kernels, como funciones polinomiales o RBF, que permiten capturar relaciones no lineales sin necesidad de crear manualmente nuevos predictores Tian y Yu (2012). Otra fortaleza clave es que la solución SVM depende de unos pocos ejemplos denominados vectores soporte, lo que facilita cierto grado de interpretabilidad: se puede identificar qué empresas determinan la frontera de decisión.

Estudios de benchmarking en la década de 2010, como el de Tsai et al. (2014), han mostrado que la SVM ofrece un rendimiento competitivo, generalmente solo superada por ensamblajes de modelos (ensembles) como XGBoost. H. Zhang et al. (2016) incluso propusieron versiones avanzadas de SVM con kernels múltiples para capturar interacciones complejas en los datos financieros de empresas chinas. Sin embargo, estudios recientes como Barboza et al. (2021) y W.-C. Chen et al. (2020) indican que aunque la SVM sigue siendo sólida, ha sido parcialmente desplazada por modelos más escalables y

versátiles.

En cuanto a limitaciones, las SVM enfrentan varios retos en entornos financieros longitudinales y desbalanceados. Primero, una SVM estándar no incorpora la secuencia temporal de manera explícita: requiere inputs estáticos o ingenierizados que codifiquen la información temporal, y no tiene memoria de trayectoria. La alternativa sería diseñar kernels específicos para series temporales, como DTW o kernels dinámicos, los cuales no son comunes en la práctica. Segundo, las SVM escalan mal a conjuntos de datos grandes: el entrenamiento puede tener una complejidad cuadrática respecto al número de observaciones, haciéndolas menos prácticas en escenarios de big data. Tercero, la SVM es altamente sensible a su calibración: la selección de hiperparámetros como el costo C y los parámetros del kernel es fundamental, y una mala configuración puede deteriorar el rendimiento o hacerla ineficiente computacionalmente. Además, ante conjuntos de datos desbalanceados, una SVM simple puede sesgarse hacia la clase mayoritaria si no se utilizan mecanismos coste-sensibles G. Wang et al. (2014).

Modelos de Aprendizaje Profundo: Redes Neuronales, LSTM y CNN

Las técnicas de aprendizaje profundo han ganado terreno en la literatura reciente de predicción de quiebras, particularmente por su capacidad de explotar patrones complejos y series temporales extensas. Este grupo incluye las redes neuronales artificiales clásicas (ANN), las redes LSTM y las redes convolucionales (CNN), cada una con particularidades útiles. Las ANN, introducidas en este campo desde los años noventa, son capaces de modelar relaciones no lineales complejas entre ratios financieros y la probabilidad de quiebra, sin requerir supuestos distribucionales estrictos Odom y Sharda (1990) y Wilson y Sharda (1994). Sin embargo, presentan desventajas como la necesidad de abundantes datos y su limitada interpretabilidad Marso et al. (2020). Las redes LSTM han sido destacadas por su habilidad para capturar dependencias temporales de largo plazo en secuencias contables. Kim et al. (2022) mostraron que, usando datos mensuales de más de una década, las LSTM superaron a modelos como SVM y Random Forest. Pellegrino et al. (2024) propusieron una arquitectura de múltiples cabezas donde cada ratio se procesa secuencialmente, logrando mayor recall y menor tasa de falsos negativos. Las CNN han sido aplicadas exitosamente de dos formas: transformando vectores de ratios en imágenes Hosaka (2019) y como redes 1D

aplicadas directamente a series temporales. Las CNN pueden detectar patrones locales (como caídas abruptas o recuperaciones transitorias), y cuando se combinan con LSTM en arquitecturas híbridas CNN-LSTM, permiten capturar tanto patrones locales como globales Qu et al. (2019). En contraste, el modelo *k-NN* funcional se basa en la comparación explícita de trayectorias financieras utilizando una métrica definida a priori. A diferencia de las redes profundas, no aprende representaciones internas optimizadas, pero ofrece interpretabilidad basada en la similitud con casos históricos reales. Este enfoque puede resultar más robusto en muestras pequeñas, mientras que las redes profundas tienden a requerir grandes volúmenes de datos Aljawazneh et al. (2021). En suma, las redes profundas ofrecen una alta capacidad predictiva a costa de mayor complejidad y opacidad, mientras que el *k-NN* funcional permite decisiones explicables basadas en analogías, siendo útil especialmente cuando el volumen de datos es limitado o se prioriza la trazabilidad del modelo.

Modelos basados en XGBoost

Uno de los algoritmos de aprendizaje automático que ha cobrado mayor relevancia en la predicción de quiebras empresariales en los últimos años es el *Extreme Gradient Boosting* (XGBoost), una implementación eficiente del algoritmo de *gradient boosting* desarrollado por Chen y Guestrin en 2016. Su capacidad para manejar datos tabulares, modelar relaciones no lineales complejas, y resistir sobreajuste mediante técnicas de regularización lo ha convertido en una opción predilecta para múltiples estudios en finanzas y riesgo empresarial.

Zieba et al. (2016) fueron de los primeros en demostrar la eficacia de XGBoost en este dominio, aplicándolo a empresas polacas con variables generadas sintéticamente para mejorar la representación de empresas quebradas. Obtuvieron mejoras sustanciales respecto a SVM y redes neuronales. De forma similar, M. Carmona et al. (2019) aplicaron XGBoost al sector bancario de Estados Unidos, mostrando que superaba consistentemente a modelos tradicionales en métricas de AUC y precisión, lo que sugiere su idoneidad incluso en entornos regulados. En Europa, Climent et al. (2019) implementaron XGBoost para anticipar crisis bancarias en la Eurozona, destacando su habilidad para identificar señales tempranas de inestabilidad financiera, reforzada por la importancia que el modelo asignaba a ratios de liquidez y apalancamiento.

Otros estudios han resaltado no solo el rendimiento de XGBoost, sino

también su capacidad de integración en sistemas automatizados. Papík y Papíková (2024b) exploraron su uso en flujos de trabajo de *AutoML*, mostrando que puede ser optimizado automáticamente sin intervención humana, logrando altos niveles de precisión y robustez en manufactura. Esto lo convierte en un candidato atractivo para sistemas de monitoreo continuo. Qian et al. (2022) abordaron el problema desde el preprocesamiento, proponiendo un nuevo método de selección de características ajustado a XGBoost que redujo la redundancia y mejoró la capacidad predictiva general, especialmente en muestras desbalanceadas.

Desde una perspectiva comparativa, Huang y Yen (2019) evaluaron exhaustivamente más de una docena de modelos de predicción financiera, concluyendo que XGBoost se ubicaba consistentemente entre los mejores en términos de F1-score y precisión, manteniendo al mismo tiempo tiempos de entrenamiento razonables. Por su parte, Son et al. (2019) utilizaron XGBoost en un enfoque híbrido de análisis de datos para predicción de bancarrota, logrando altos niveles de exactitud incluso sin requerir complejos ajustes paramétricos.

El desempeño técnico de XGBoost también ha sido explorado en combinación con técnicas de explicabilidad. Park et al. (2021) analizaron distintos modelos de predicción bajo criterios de interpretabilidad, mostrando que XGBoost, aunque menos transparente que modelos lineales, podía ser interpretado parcialmente usando herramientas como SHAP para identificar los factores financieros más influyentes en cada predicción. Esta línea es crucial para entornos regulados donde la trazabilidad del modelo es indispensable.

En cuanto a estrategias de validación, du Jardin (2016) propusieron una técnica de clasificación en dos etapas que utiliza XGBoost como componente clave, con resultados que sugieren una mejor discriminación de empresas en riesgo. Finalmente, estudios como Aljawazneh et al. (2021) han corroborado el rendimiento de XGBoost en comparación con arquitecturas profundas de aprendizaje, destacando su eficiencia computacional, su facilidad de entrenamiento y su solidez en conjuntos de datos ruidosos o escasos.

En conjunto, la literatura reciente posiciona a XGBoost como un estándar de referencia en tareas de predicción de quiebra. Sus ventajas combinan precisión, escalabilidad y adaptabilidad a distintos contextos empresariales, industriales y geográficos, lo cual lo convierte en un punto de comparación natural para cualquier modelo novedoso, incluido el enfoque funcional basado en k-NN que se propone en esta tesis.

Modelos Híbridos

En esta línea, varios autores han mostrado que combinar modelos heterogéneos mediante un meta-modelo mejora el desempeño. Por ejemplo, H. Zhang et al. (2016) proponen un enfoque basado en subespacios no lineales y kernels múltiples para capturar diferentes aspectos de las trayectorias financieras, mientras que Tsai et al. (2014) comparan múltiples ensamblajes, incluyendo combinaciones de SVM, redes neuronales y árboles de decisión, destacando la superioridad del stacking en términos de exactitud promedio. Ji et al. (2025) desarrollan un sistema de predicción basado en Gradient Boosting que incorpora un flujo dinámico para adaptarse a cambios financieros en el tiempo. Estos trabajos muestran que el uso de ensembles no solo mejora el rendimiento, sino que también mitiga la varianza de los modelos individuales.

Varios estudios han explorado la hibridación entre redes neuronales y algoritmos genéticos para optimizar pesos, topologías o variables de entrada. Ansari et al. (2020) presentan una red neuronal entrenada con un método metaheurístico híbrido (GA y BSO), mostrando mejoras notables en velocidad de convergencia y precisión. De forma similar, Ansah-Narh et al. (2024) proponen una arquitectura que combina adaptación de dominio con GA para seleccionar atributos relevantes, alcanzando altos niveles de generalización. Este tipo de hibridación también ha sido documentado por F. Lin et al. (2011), quienes integran aprendizaje de manifolds y SVM, logrando reducir el error tipo II. Estas estrategias son especialmente útiles para escapar de mínimos locales y mejorar la estabilidad del aprendizaje.

Aljawazneh et al. (2021) comparan redes LSTM, GRU y bidireccionales para capturar dinámicas temporales, encontrando que los modelos híbridos que integran atención o estructuras bidireccionales superan a las redes simples. Feng et al. (2019) combinan redes neuronales con embeddings textuales de informes financieros, mientras que Hosaka (2019) emplean CNN sobre imágenes generadas a partir de ratios contables. Estas arquitecturas reflejan una tendencia a enriquecer el input con datos no estructurados, combinando visión computacional y NLP con ratios financieros.

G. Wang et al. (2014) proponen un modelo boosting sensible al costo que incorpora selección de características, mientras que W.-C. Chen et al. (2020) combinan boosting y SMOTE para tratar el desbalance de clases. Por su parte, Y.-M. Lin y Chang (2023) desarrollan una arquitectura CNN con explicabilidad basada en SHAP, lo que permite visualizar qué variables motivan una predicción específica. Esta tendencia es reforzada por Zhao

et al. (2024a), quienes integran análisis de redes complejas para representar las relaciones interempresariales, y mejoran la interpretabilidad a través de visualizaciones de centralidad y vecindad.

Capítulo 3

Marco metodológico: diseño del modelo funcional y métrica personalizada

Este capítulo describe el diseño metodológico del modelo propuesto para predecir el riesgo de quiebra empresarial a partir de trayectorias financieras. El enfoque central consiste en representar a cada empresa como un objeto funcional multivariado, permitiendo comparar su evolución en el tiempo mediante una métrica de distancia personalizada. Este marco funcional no solo respeta la naturaleza secuencial de los datos contables, sino que también incorpora mecanismos de penalización ante trayectorias truncadas, diferencias de escala y valores extremos.

Aunque históricamente la predicción de quiebra se ha basado en enfoques estáticos —que utilizan únicamente la fotografía financiera de un año específico—, en los últimos años se han desarrollado métodos dinámicos que incorporan la dimensión temporal, como modelos de *boosting* con variables rezagadas o redes neuronales recurrentes tipo LSTM (Kuizinienė et al., 2022; Nazareth & Reddy, 2023). Sin embargo, estos modelos no comparan directamente trayectorias entre empresas, sino que las transforman en secuencias o vectores tabulares, perdiendo parte de la estructura funcional del problema.

En este estudio se avanza un paso más, proponiendo una representación funcional explícita de las trayectorias financieras, que permite capturar de manera más estructurada su evolución a lo largo del tiempo. Este tipo de enfoque ha sido ampliamente estudiado bajo el marco del análisis funcional de datos (FDA), cada vez más presente en campos como medicina, climatología

o finanzas (Almanjahie et al., 2024).

La propuesta metodológica incluye, además, mecanismos para ampliar el espacio funcional original mediante la incorporación de variables categóricas y cuantitativas, estáticas o secuenciales.

El capítulo se organiza de la siguiente forma. En primer lugar, se justifica conceptualmente el uso del enfoque funcional en contextos de riesgo financiero empresarial. Luego, se construye formalmente el espacio funcional $\mathcal{F}$ mediante ventanas móviles temporales. A continuación, se define la métrica funcional penalizada propuesta, detallando cada una de sus transformaciones y su expresión computacional discreta. Finalmente, se presentan ejemplos gráficos, variantes metodológicas alternativas, y se discuten extensiones generales del enfoque que permiten incorporar variables mixtas y adaptar el modelo a estructuras empresariales.

3.1. Construcción del Espacio Funcional

El enfoque funcional adoptado en esta investigación representa a cada empresa como una trayectoria multivariada de indicadores financieros a lo largo del tiempo. Esta representación no se construye a partir de toda la historia contable de la empresa, sino que se delimita mediante una ventana temporal de longitud fija, aplicada retrospectivamente desde el último año con información reportada por cada entidad.

Esta decisión metodológica surge de una reflexión inspirada en modelos secuenciales como las redes neuronales tipo LSTM, que procesan información en forma de secuencias ordenadas. Aunque este estudio no recurre a tales arquitecturas, el razonamiento que las sustenta —basado en la captura de trayectorias estructuradas— motivó la adopción de una estrategia similar, orientada a preservar el componente dinámico en la evolución financiera de las empresas.

Inicialmente se consideró utilizar toda la información disponible para cada empresa, desde su primer hasta su último año registrado. Sin embargo, esta aproximación condujo a una alta heterogeneidad en la longitud de las trayectorias y a una proporción significativa de valores faltantes, lo cual habría requerido aplicar técnicas de imputación extensiva. Dichas técnicas, si bien útiles en algunos contextos, podían introducir patrones espurios o sesgos que comprometieran la validez del análisis.

Una alternativa habría sido tomar únicamente las empresas activas en un año fijo (por ejemplo, el último año del periodo de estudio) y construir trayectorias hacia atrás desde ese punto. No obstante, esta estrategia implicaría excluir todas las empresas que dejaron de reportar antes de ese año, muchas de las cuales son precisamente aquellas que enfrentaron deterioro financiero o procesos de liquidación. Esto habría generado un sesgo de selección relevante, limitando la posibilidad de aprender de las trayectorias que efectivamente condujeron a la quiebra.

Frente a estas limitaciones, se optó por una solución intermedia y más robusta: fijar una longitud ℓ común para todas las trayectorias y construir, para cada empresa, una ventana temporal de ℓ años consecutivos, anclada al último año para el cual se tiene información disponible. Esta estrategia permite estandarizar la representación temporal sin perder observaciones valiosas ni depender en exceso de imputaciones, garantizando así la inclusión tanto de empresas vigentes como de aquellas que eventualmente salieron del sistema.

El valor de esta estrategia ha sido respaldado también en estudios recientes. Por ejemplo, Abrahamsen et al. (2024) emplean modelos de *boosting* como LightGBM sobre datos financieros trimestrales de empresas nórdicas, estructurados mediante una *ventana móvil* para capturar la evolución dinámica del riesgo de crédito. Sus resultados muestran que este tipo de representación temporal mejora significativamente la precisión predictiva frente a enfoques estáticos, lo que valida empíricamente la pertinencia de estructurar trayectorias multivariadas en contextos de predicción de insolvencia.

Formalización de la representación funcional.

Sea ℓ la longitud de la ventana temporal, expresada en años consecutivos. Para garantizar la comparabilidad entre empresas que reportaron en distintos periodos, se define una escala de tiempo relativa funcional, denotada por $\tau \in \{-\ell + 1, -\ell + 2, \ldots, 0\}$. En esta escala:

- $\tau = 0$ corresponde al último año con información disponible para la empresa,
- Los valores negativos representan los años anteriores dentro de la misma ventana,
- Esta notación permite comparar trayectorias sin hacer referencia al año calendario real.

Cada empresa e es representada por un vector funcional multivariado:

$$\mathbf{X}^e(\tau) = \begin{bmatrix} X_1^e(\tau) \\ X_2^e(\tau) \\ \vdots \\ X_m^e(\tau) \end{bmatrix}, \quad \tau \in \{-\ell + 1, \ldots, 0\}$$

donde $X_j^e(\tau)$ representa el valor observado del indicador financiero j en el año relativo τ, dentro de la trayectoria de la empresa.

Definición del espacio funcional. El espacio funcional se define como:

$$\mathcal{F} \subset \{\mathbf{X} : [-\ell + 1, 0] \to \mathbb{R}^m\}$$

donde cada $\mathbf{X} \in \mathcal{F}$ es una función multivariada que asocia a cada instante τ un vector de m indicadores financieros.

No se impone ninguna condición adicional de continuidad ni integrabilidad sobre estas funciones, ya que la métrica funcional propuesta operará directamente sobre evaluaciones puntuales de las trayectorias. La robustez frente a valores extremos o faltantes se incorpora mediante mecanismos explícitos de penalización y acotamiento, sin requerir supuestos fuertes sobre la forma funcional global.

Cada elemento de $\mathcal{F}$ es una función vectorial definida sobre un intervalo temporal relativo de longitud ℓ, que describe la evolución conjunta de m indicadores financieros para una empresa dada. El dominio está formado por el intervalo continuo $[-\ell + 1, 0]$, mientras que el codominio corresponde al espacio $\mathbb{R}^m$, es decir, el conjunto de vectores contables que caracterizan la situación financiera de la empresa en cada instante τ.

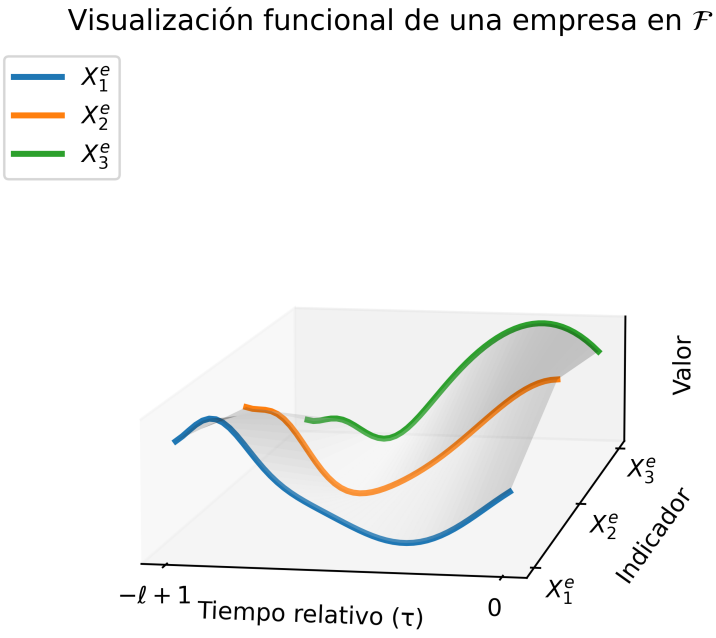

Figura 3.1: Ejemplo esquemático de un elemento del espacio funcional $\mathcal{F}$. Cada curva representa la evolución de un indicador financiero a lo largo del tiempo relativo $\tau \in [-\ell + 1, 0]$.
Fuente: Elaboración propia.

Para que una empresa e pertenezca al espacio funcional $\mathcal{F}$, debe cumplir con las siguientes condiciones mínimas:

- Disponer de información financiera correspondiente a un intervalo temporal de longitud ℓ, con cobertura suficiente de los indicadores requeridos,
- No presentar valores estructuralmente ausentes en la totalidad de algún indicador durante el intervalo considerado,

Este espacio funcional $\mathcal{F}$ constituye el dominio sobre el cual se definirá la métrica de similitud entre empresas propuesta en la siguiente sección.

3.2. Definición de la métrica funcional personalizada

A continuación, se presenta la construcción formal de la métrica funcional utilizada para comparar empresas representadas como elementos en $\mathcal{F}$. La métrica se define en cuatro etapas: cálculo de distancias por indicador, penalización de valores faltantes, acotamiento de extremos y combinación ponderada.

Paso 1: Distancia funcional acumulada por indicador.

Sea $X_j^e(\tau)$ la trayectoria funcional del indicador financiero j para la empresa e, definida sobre el intervalo relativo $[-\ell+1, 0]$. Para comparar dos trayectorias $X_j^e(\tau)$ y $X_j^{e'}(\tau)$, se considera el mayor subintervalo compartido $[k, 0] \subseteq [-\ell+1, 0]$, donde ambas funciones están definidas.

La distancia funcional acumulada (no penalizada) se define como:

$$d_j^{\text{acum}}(X_j^e, X_j^{e'}) = \int_k^0 \left| X_j^e(\tau) - X_j^{e'}(\tau) \right| d\tau$$

esta forma permite acumular de forma continua las diferencias puntuales entre indicadores financieros a lo largo del tiempo relativo. De este modo, se mide no solo la magnitud de la diferencia de los indicadores tiempo a tiempo, sino también su persistencia temporal, respetando la estructura secuencial de la trayectoria.

Nota: Si ambas trayectorias están completamente definidas, es decir, si $k = -\ell + 1$, la distancia se calcula directamente sobre todo el intervalo:

$$d_j^{\text{acum}}(X_j^e, X_j^{e'}) = \int_{-\ell+1}^0 \left| X_j^e(\tau) - X_j^{e'}(\tau) \right| d\tau$$

La Figura 3.2 muestra un ejemplo ilustrativo con tres indicadores. En el caso del CTN, ambas trayectorias están completamente definidas desde $-\ell + 1$, por lo que la distancia se calcula de forma integral sobre todo el intervalo. Para ROA, la trayectoria de la empresa e_2 comienza en un punto intermedio k_1, por lo que la distancia solo se acumula a partir de ese punto, situación que se considerará más adelante en la penalización. En el caso de la Liquidez Corriente, la trayectoria de e_2 presenta un comportamiento extremo que tiende a infinito cerca de $\tau = 0$, generando una distancia acumulada divergente que será posteriormente acotada en el paso correspondiente.

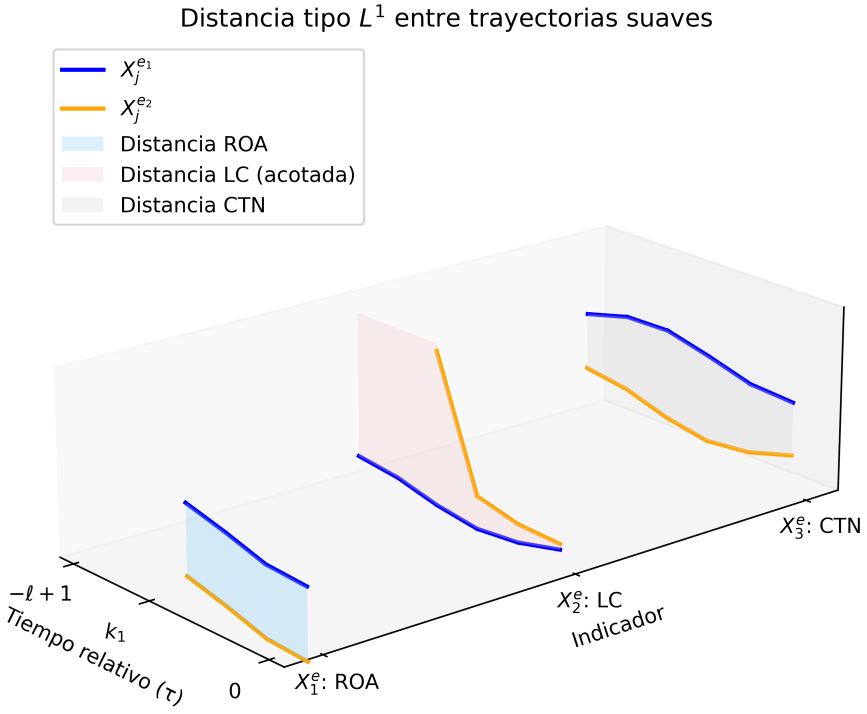

Figura 3.2: Ejemplo visual del cálculo de la distancia funcional acumulada tipo L^1 entre trayectorias suaves de dos empresas.

Paso 2: Penalización por pérdida de dominio.

Cuando el intervalo de comparación $[k, 0]$ no cubre la totalidad de la ventana funcional, se aplica una penalización proporcional a la fracción de trayectoria no utilizada:

$$d_j(X_j^e, X_j^{e'}) = d_j^{\text{acum}}(X_j^e, X_j^{e'}) \cdot \left(1 + \lambda \cdot \frac{k + \ell - 1}{\ell}\right)$$

donde $\lambda \in \mathbb{R}^+$ es un parámetro de penalización por pérdida de información y $\frac{k+\ell-1}{\ell}$ es la proporción de la trayectoria no disponible.

Este factor penaliza proporcionalmente la fracción de trayectoria no disponible, sin alterar la estructura de acumulación puntual sobre los datos

efectivamente observados. Además, dado que la pertenencia al espacio funcional $\mathcal{F}$ exige que toda empresa tenga información disponible en $\tau = 0$, se garantiza que siempre existe al menos un rango de comparación para cualquier par de empresas.

Nota técnica. La función de distancia aquí propuesta puede no cumplir estrictamente con la desigualdad triangular, debido a la penalización aplicada por pérdida de información. Por tanto, debe entenderse como una *semi-métrica funcional penalizada*, válida en contextos donde el objetivo no es construir una topología matemática formal, sino capturar relaciones de similitud entre trayectorias potencialmente incompletas. Esta aproximación es consistente con desarrollos recientes en finanzas aplicadas, donde el uso de semi-métricas ha sido justificado como una estrategia efectiva para comparar estructuras temporales complejas, incluso en presencia de choques estructurales o valores extremos N. James et al., 2023.

La Figura 3.3 muestra cómo se representa gráficamente la penalización por pérdida de dominio en un único indicador (ROA), cuando una de las trayectorias no tiene datos en los dos primeros años. La penalización se ilustra como un área adicional (en verde claro) que se agrega desde el piso, proporcional a la fracción de la trayectoria no disponible.

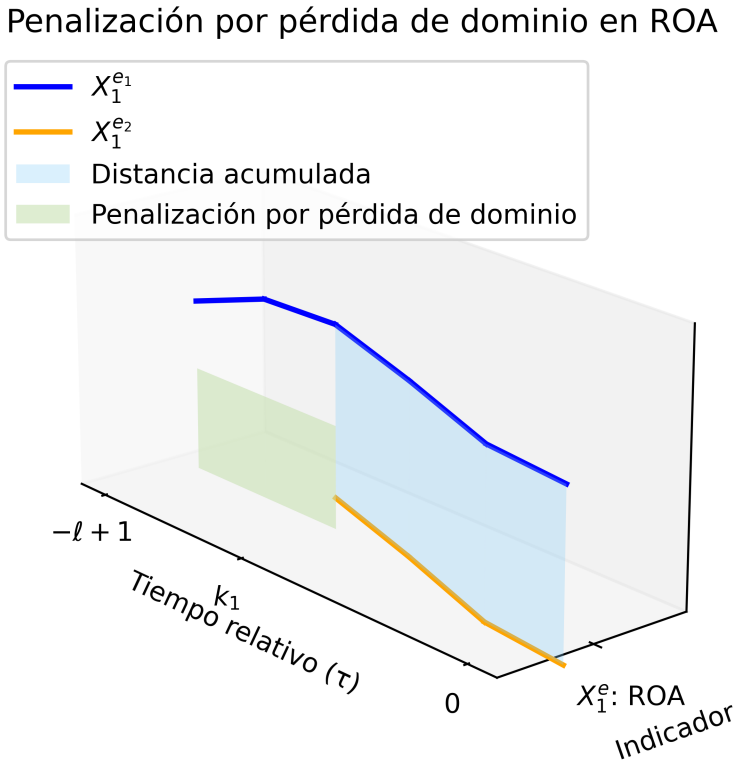

Figura 3.3: Penalización funcional aplicada a la trayectoria del indicador ROA. La franja verde representa la penalización agregada cuando no hay datos en los dos primeros años.

La Figura 3.4 muestra el mismo ejemplo en el contexto completo de los tres indicadores considerados. Solo ROA presenta pérdida de dominio, por lo que es el único que incorpora la penalización proporcional. Los demás indicadores mantienen su distancia acumulada original.

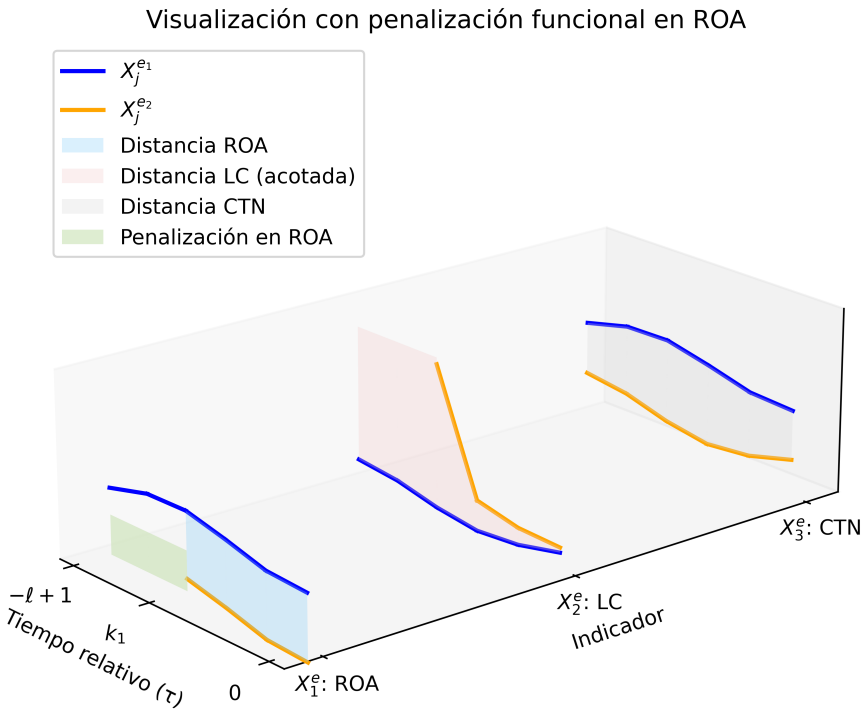

Figura 3.4: Visualización conjunta con los tres indicadores. Solo ROA presenta penalización por pérdida de dominio.

Paso 3: Acotamiento de extremos.

En ciertos indicadores financieros, como la liquidez corriente, es posible que se presenten valores infinitos cuando el denominador (por ejemplo, el pasivo corriente) es igual a cero y el numerador es positivo. Estos casos no corresponden a errores de digitación, sino a situaciones contables reales que deben ser consideradas con cuidado. Para evitar que este tipo de valores extremos distorsione el cálculo de distancias, se aplica una transformación que acota la distancia máxima de cada indicador al intervalo $[0, 1)$, sin necesidad de imputar ni modificar los valores originales:

$$\tilde{d}_j(X_j^e, X_j^{e'}) = \frac{d_j(X_j^e, X_j^{e'})}{1 + d_j(X_j^e, X_j^{e'})}$$

Esta transformación suaviza las diferencias extremas sin alterar el orden relativo de las trayectorias y sin forzar una imputación arbitraria de valores fuera de escala.

Es importante aclarar que, en esta metodología, no se realiza ninguna normalización previa sobre las trayectorias financieras. A diferencia de los enfoques tradicionales de aprendizaje automático —que suelen transformar las variables de entrada mediante escalamiento Min-Max o estandarización Z-score—, aquí se conserva la escala original de cada indicador.

Esta decisión metodológica responde a dos razones fundamentales. Primero, se busca preservar la interpretabilidad económica directa de las trayectorias: por ejemplo, una diferencia de 5 unidades en el ROA o en el Capital de Trabajo mantiene su significado contable. Segundo, la métrica funcional propuesta incorpora mecanismos internos que permiten controlar el efecto de escalas heterogéneas y valores extremos. En particular, la penalización proporcional por pérdida de dominio y la transformación acotadora de tipo racional ($\tilde{d}_j = \frac{d_j^{(pen)}}{1+d_j^{(pen)}}$) aseguran que ninguna dimensión pueda dominar de manera desproporcionada el cálculo de la distancia total, incluso si contiene valores infinitos o altamente desbalanceados.

En este enfoque, por tanto, la "normalización.ºcurre a nivel de salida y no de entrada. Además, la posibilidad de asignar pesos específicos w_j en la combinación final de distancias brinda un control adicional sobre la importancia relativa de cada indicador, el cual será optimizado posteriormente mediante técnicas de búsqueda como Optuna.

Nota técnica. La transformación racional $\tilde{d}(x, y) = \frac{d(x,y)}{1+d(x,y)}$ es continua, estrictamente creciente y acota la distancia al intervalo $[0, 1)$ en términos matemáticos. No obstante, en esta implementación, cuando la distancia penalizada $d(x, y)$ tiende a infinito, se asigna directamente $\tilde{d}(x, y) = 1$ como valor máximo por convención práctica. Esta decisión permite representar explícitamente casos extremos sin perder la estructura relativa de las distancias. Si bien en contextos donde d es una métrica formal la transformación mantiene propiedades topológicas equivalentes, aquí se aplica sobre una semi-métrica penalizada, adecuada para comparar trayectorias incompletas en contextos financieros Rudin, 1976.

A continuación se ilustra gráficamente la transformación acotada $\tilde{d}_j = \frac{d_j}{1+d_j}$, aplicada a tres indicadores financieros (ROA, LC y CTN). En la Figura 3.5, cada curva representa cómo se transforma la distancia funcional acumula-

da original d_j en su versión acotada $\tilde{d}_j$, para cada uno de los indicadores. Los puntos de color marcan ejemplos concretos de distancias acumuladas, donde se observa que incluso si $d_j \to \infty$, la distancia acotada se aproxima asintóticamente a 1, como ocurre con el caso de la liquidez corriente (LC).

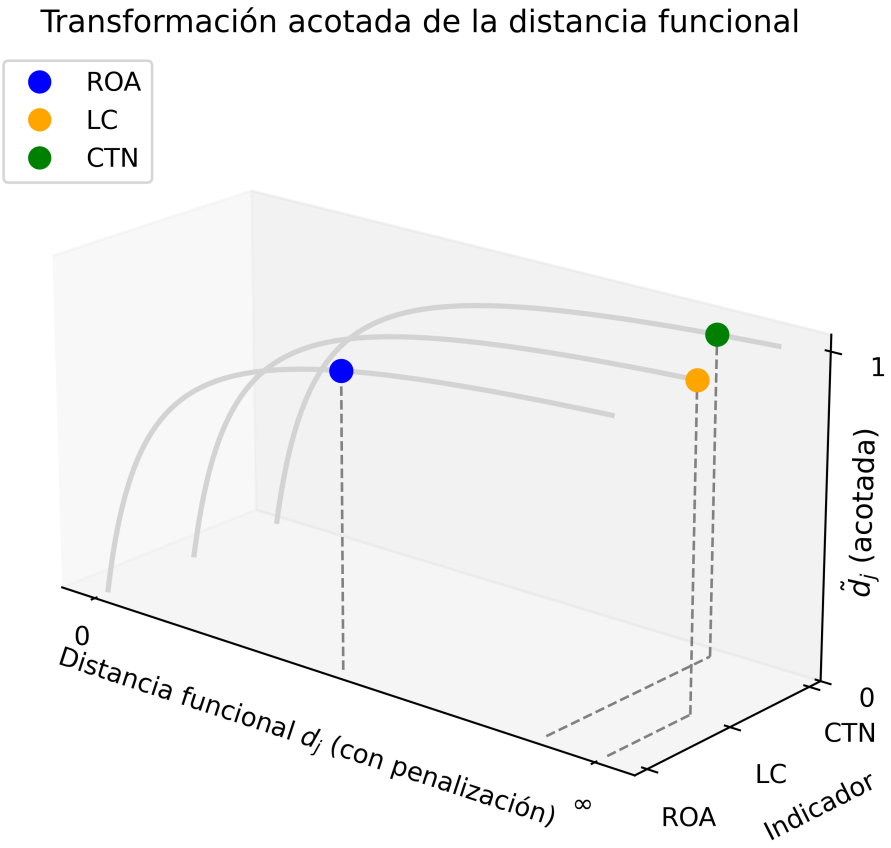

Figura 3.5: Transformación acotada de la distancia funcional. Cada curva muestra la relación entre d_j y $\tilde{d}_j$ para los indicadores ROA, LC y CTN. Los puntos de colores representan casos específicos: ROA con una distancia moderada, CTN con una distancia alta y LC con un valor extremo tendiente a infinito, acotado a $\tilde{d}_j = 1$.

Paso 4: Combinación ponderada de indicadores.

Una vez calculadas las distancias acotadas $\tilde{d}_j(X_j^e, X_j^{e'})$ para cada uno de los m indicadores financieros, se obtiene la distancia funcional total entre dos empresas mediante un promedio ponderado:

$$D(X^e, X^{e'}) = \sum_{j=1}^{m} w_j \cdot \tilde{d}_j(X_j^e, X_j^{e'})$$

donde $w_j \in [0, 1]$ representa el peso asignado al indicador j, con la restricción $\sum_{j=1}^{m} w_j = 1$.

Esta estructura permite ajustar la importancia relativa de cada indicador en función de criterios técnicos, regulatorios o empíricos. Los pesos w_j pueden asignarse de forma uniforme o ser optimizados con base en la capacidad predictiva de cada indicador, mediante técnicas como validación cruzada o búsqueda bayesiana.

Nota técnica. La distancia funcional total $D(X^e, X^{e'})$ se construye como una combinación ponderada de métricas individuales $\tilde{d}_j$. Esta estructura conserva las propiedades métricas clásicas. En particular, la desigualdad triangular se preserva gracias a la linealidad de la suma:

$$D(x, z) = \sum_j w_j \cdot d_j(x, z) \leqslant \sum_j w_j \cdot [d_j(x, y) + d_j(y, z)] = D(x, y) + D(y, z)$$

Esta propiedad está documentada en la literatura matemática clásica sobre espacios métricos y se discute formalmente en Ó Searcóid, 2007.

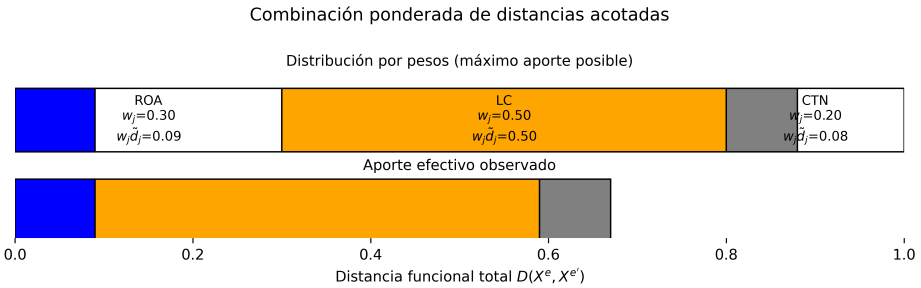

Figura 3.6: Combinación ponderada de distancias acotadas.

La figura ilustra gráficamente cómo se construye la distancia funcional total $D(X^e, X^{e'})$ a partir de las distancias acotadas $\tilde{d}_j$ y los pesos w_j asignados a cada indicador.

La barra superior representa la estructura de ponderación global: cada sección coloreada muestra el máximo aporte que podría tener un indicador si su distancia fuera igual a 1. En cambio, la barra inferior refleja el aporte real de cada indicador para un par específico de empresas. Este contraste permite visualizar en qué medida cada indicador efectivamente contribuye a la distancia total, más allá del peso que le haya sido asignado inicialmente.

Expresión explícita de la métrica funcional completa.
Para efectos de claridad y referencia técnica, a continuación se presenta la expresión total que integra las cuatro capas de la métrica funcional desarrollada:

$$D(X^e, X^{e'}) = \sum_{j=1}^{m} w_j \left(\frac{\int_{k_j}^{0} \left| X_j^e(\tau) - X_j^{e'}(\tau) \right| d\tau \cdot \left(1 + \lambda \frac{k_j + \ell - 1}{\ell}\right)}{1 + \int_{k_j}^{0} \left| X_j^e(\tau) - X_j^{e'}(\tau) \right| d\tau \cdot \left(1 + \lambda \frac{k_j + \ell - 1}{\ell}\right)} \right) \quad (3.2.1)$$

Donde:

- e y e' representan dos empresas distintas.
- $X_j^e(\tau)$ es la trayectoria temporal del indicador financiero j para la empresa e, evaluada sobre el tiempo relativo $\tau \in [-\ell + 1, 0]$.
- ℓ es la longitud total de la ventana funcional, expresada en años consecutivos.
- $k_j \in [-\ell + 1, 0]$ es el mayor valor tal que ambas trayectorias X_j^e y $X_j^{e'}$ están definidas sobre el subintervalo $[k_j, 0]$.
- $\lambda \in \mathbb{R}^+$ es el parámetro de penalización por pérdida de información.
- $w_j \in [0, 1]$ es el peso asignado al indicador j, con $\sum_{j=1}^{m} w_j = 1$.

Dicha expresión destaca que la métrica propuesta no puede ser considerada como una simple instancia de una distancia L^1, sino como una construcción compuesta con propiedades funcionales específicas.

Discusión metodológica y consideraciones adicionales sobre la métrica

La métrica funcional propuesta en este trabajo presenta una serie de ventajas sustantivas desde el punto de vista conceptual, técnico y aplicado. En primer lugar, permite representar la evolución financiera de cada empresa como una trayectoria multivariada definida sobre una escala temporal relativa, lo cual facilita la comparación entre firmas sin necesidad de alinear los datos por año calendario. Esta decisión metodológica se fundamenta en que lo relevante no es el momento histórico en el que ocurre un evento de quiebra, sino la forma en que se comportan los indicadores contables en los años previos al desenlace observado.

Desde esta perspectiva, el uso de una ventana móvil retrospectiva —por ejemplo, de los últimos cinco años reportados— constituye una estrategia que deliberadamente abstrae el análisis de factores macroeconómicos coyunturales, como el COVID-19, las crisis sectoriales o los cambios en las tasas de interés. A diferencia de otros enfoques que anclan las observaciones al calendario (comparando empresas en un mismo año), la presente propuesta se basa en un sistema de tiempo relativo centrado en la trayectoria individual de cada empresa. Esto permite comparar estructuras evolutivas equivalentes, independientemente del momento histórico en el que se produjeron.

Este diseño ofrece dos beneficios clave. Primero, evita que fenómenos externos afecten la comparación funcional entre empresas que atravesaron condiciones macroeconómicas distintas pero exhibieron trayectorias contables similares. Segundo, habilita el uso de información proveniente de empresas ya inactivas —como aquellas que quebraron hace más de una década— las cuales quedarían excluidas de los análisis si se exigiera sincronía temporal. En consecuencia, la métrica funcional no se ve sesgada por sincronías espurias y se concentra en lo que interesa desde una perspectiva de gestión: la evolución interna y progresiva de los indicadores financieros, independientemente de los eventos exógenos del entorno.

En este sentido, el enfoque funcional se aproxima conceptualmente a modelos dinámicos como LSTM, que también capturan trayectorias a lo largo del tiempo, pero preserva la interpretabilidad y la trazabilidad de cada componente del cálculo, sin recurrir a arquitecturas de caja negra.

Adicionalmente, la métrica incorpora mecanismos explícitos de penalización proporcional por pérdida de información, lo que evita la imputación arbitraria de valores faltantes y permite conservar los datos observados. Tam-

bién se incluye una transformación acotada para evitar que indicadores con valores extremos (como la liquidez corriente infinita) dominen la distancia total. El uso de pesos por variable —ajustables mediante validación cruzada— añade un nivel adicional de interpretabilidad, ya que permite identificar la contribución relativa de cada indicador financiero a la distancia entre empresas. Esta estructura convierte al modelo k-NN funcional, tradicionalmente no paramétrico y opaco, en una herramienta transparente y adaptable para tareas de predicción y análisis explicativo.

Cabe señalar que, aunque este enfoque no se fundamenta en una estructura probabilística formal (como ocurre con los modelos de supervivencia o las redes LSTM bayesianas), la proporción de vecinos con estado crítico puede interpretarse como un score empírico de vulnerabilidad. Esto habilita el uso de métricas de evaluación como el AUC-ROC, sin necesidad de estimar funciones de riesgo o likelihoods.

No obstante, esta estrategia también presenta limitaciones. Al utilizar tiempo relativo, se pierde sincronía con fenómenos macroeconómicos definidos en el calendario, como políticas públicas o crisis sectoriales. Además, si existen variables contables altamente correlacionadas, su peso combinado podría inducir redundancia en la métrica si no se realiza un análisis previo de colinealidad. Aunque la colinealidad no representa un problema en modelos k-NN con métricas simétricas clásicas, sí puede generar sesgos si se asignan pesos diferenciados a indicadores redundantes, como ocurre en esta propuesta. Por tanto, se recomienda realizar un análisis exploratorio previo o considerar herramientas como el análisis de componentes principales (PCA) para reducir redundancias antes de la estimación de pesos.

Desde el punto de vista operativo, la métrica funcional implica un mayor costo computacional que las métricas tradicionales aplicadas sobre vectores planos, especialmente si se utiliza validación cruzada o búsqueda bayesiana de hiperparámetros. Asimismo, la longitud de la ventana funcional ℓ afecta la cantidad de datos disponibles y la sensibilidad del modelo: valores muy pequeños pueden no capturar bien la evolución, mientras que valores muy grandes aumentan los casos con datos faltantes. En este trabajo se ha utilizado un valor fijo, pero es factible considerar a ℓ como un hiperparámetro adicional y optimizarlo mediante validación cruzada, con base en métricas como el F1-score promedio o su desviación.

Finalmente, el diseño modular de la métrica permite explorar variantes estructurales que podrían servir como ejercicios de sensibilidad metodológica: la eliminación de la penalización para conservar la propiedad métrica estricta,

el uso de ventanas fijas por año calendario, o la sustitución del acotamiento por una normalización previa de los indicadores. Estas alternativas podrían desarrollarse como extensiones futuras o comparaciones complementarias para validar la robustez del enfoque funcional propuesto.

3.3. Discretización de la métrica funcional para implementación práctica

Si bien la formulación matemática de la métrica propuesta se desarrolla en el marco del Análisis Funcional de Datos (FDA), su aplicación práctica requiere trabajar con observaciones discretas, ya que los estados financieros de las empresas son reportados en intervalos anuales. No obstante, este enfoque no implica abandonar la perspectiva funcional: en FDA aplicada es común representar funciones mediante trayectorias discretas, capturando su comportamiento evolutivo a partir de puntos relevantes sobre el dominio temporal. En consecuencia, cada empresa sigue siendo conceptualizada como una trayectoria funcional multivariada, aunque operativamente se implemente como una matriz de datos contables observados sobre una rejilla temporal finita.

Este proceso de discretización no implica abandonar el enfoque funcional. Al contrario, es una práctica común y aceptada dentro de la literatura de FDA, donde las funciones reales —por ejemplo, curvas de crecimiento, series temporales biológicas o trayectorias económicas— son representadas mediante un número finito de evaluaciones puntuales. Tal como se observa en aplicaciones en medicina, climatología o ingeniería, la información funcional suele provenir de muestras discretas evaluadas en rejillas temporales específicas.

En el caso de esta investigación, dicha rejilla temporal corresponde a la ventana móvil retrospectiva definida previamente, la cual considera los últimos ℓ años consecutivos con información disponible para cada empresa, contados hacia atrás desde su último año reportado. Esta elección permite capturar la trayectoria financiera más reciente y relevante, evitando depender del año calendario o de eventos exógenos compartidos.

Por tanto, cada trayectoria funcional multivariada $\mathbf{X}^e(\tau) \in \mathcal{F}$ es implementada como una matriz de dimensión $m \times \ell$, donde las filas representan los indicadores financieros y las columnas corresponden a posiciones relativas en el tiempo. Esta estructura preserva la esencia del enfoque funcional: el análisis se realiza sobre la forma temporal de los indicadores, respetando su secuencia y evolución interna, en lugar de tratarlos como atributos independientes o desordenados.

En las siguientes secciones se explicará cómo esta discretización se aplica a los datos reales disponibles, cómo se estructura el conjunto de observaciones, y cómo se implementa la métrica funcional sobre estas representaciones

matriciales.

Representación discreta de cada empresa

Para implementar la métrica funcional propuesta sobre datos contables reales, se adopta a partir de esta sección una notación discreta, que facilita el procesamiento computacional y la codificación estructurada de los registros. Este cambio no implica abandonar el marco funcional original, sino que responde a una práctica común en Análisis Funcional de Datos (FDA), donde las funciones son observadas a través de un conjunto finito de puntos en el dominio temporal.

Cada trayectoria funcional $X_j^e(\tau)$, correspondiente al indicador j de la empresa e, es reemplazada por un vector discreto:

$$(x_{j,t}^e)_{t=-\ell+1}^{0}$$

donde $x_{j,t}^e \in \mathbb{R}$ representa el valor observado del indicador financiero j para la empresa e en el instante relativo t, $t \in \{-\ell+1, -\ell+2, \ldots, 0\}$ define el tiempo relativo, contable hacia atrás desde el último año con información disponible.

La representación completa de la empresa e se estructura como una matriz de dimensión $m \times \ell$, donde cada fila corresponde a un indicador y cada columna a un instante temporal:

$$\mathbf{x}^e = \begin{bmatrix} x_{1,-\ell+1}^e & x_{1,-\ell+2}^e & \cdots & x_{1,0}^e \\ x_{2,-\ell+1}^e & x_{2,-\ell+2}^e & \cdots & x_{2,0}^e \\ \vdots & \vdots & & \vdots \\ x_{m,-\ell+1}^e & x_{m,-\ell+2}^e & \cdots & x_{m,0}^e \end{bmatrix} \in \mathbb{R}^{m \times \ell}$$

Esta estructura discreta permite conservar la secuencia temporal relativa y la dimensión financiera de cada empresa. La métrica funcional, tal como fue formulada en su versión continua, se aplica sobre esta representación utilizando sumas ponderadas en lugar de integrales, como se mostrará a continuación.

Métrica funcional en versión discreta

Una vez discretizadas las trayectorias funcionales, la distancia entre dos empresas e y e' se calcula sobre sus respectivas matrices $\mathbf{x}^e, \mathbf{x}^{e'} \in \mathbb{R}^{m \times \ell}$, las

cuales contienen los valores observados de cada indicador financiero en una escala temporal relativa uniforme.

La formulación teórica de la métrica, expresada en la Ecuación (3.2.1), se implementa computacionalmente mediante sumas discretas sobre los años compartidos por ambas empresas. Para cada indicador j, se calcula la diferencia acumulada sobre el subintervalo común, se aplica la penalización proporcional por pérdida de dominio, se acota el resultado para evitar dominancia por extremos, y finalmente se combinan todas las distancias usando un promedio ponderado por variable. Esta versión discreta queda expresada formalmente como:

$$D(x^e, x^{e'}) = \sum_{j=1}^{m} w_j \cdot \left(\frac{\sum_{t=k_j}^{0} \left| x_{j,t}^{e} - x_{j,t}^{e'} \right| \cdot \left(1 + \lambda \cdot \frac{k_j + \ell - 1}{\ell}\right)}{1 + \sum_{t=k_j}^{0} \left| x_{j,t}^{e} - x_{j,t}^{e'} \right| \cdot \left(1 + \lambda \cdot \frac{k_j + \ell - 1}{\ell}\right)} \right) \tag{3.3.1}$$

Donde:

- $x_{j,t}^{e} \in \mathbb{R}$ es el valor del indicador financiero j para la empresa e en el año relativo t,
- $k_j \in \{-\ell + 1, \ldots, 0\}$ es el mayor año relativo tal que ambas empresas tienen datos observados en el intervalo $[k_j, 0]$,
- $\lambda \in \mathbb{R}^+$ es el parámetro de penalización por pérdida de dominio,
- ℓ es la longitud total de la ventana temporal,
- $w_j \in [0, 1]$ es el peso del indicador j, con $\sum_{j=1}^{m} w_j = 1$.

Nota. Esta versión discreta respeta la estructura conceptual de la métrica funcional continua. Cada capa —acumulación temporal, penalización, acotamiento y combinación ponderada— se conserva íntegramente, manteniendo tanto la flexibilidad como la interpretabilidad del diseño original. *Observación.* La Ecuación (3.3.1) es matemáticamente equivalente a la formulación continua de la métrica (Ecuación (3.2.1)) cuando cada trayectoria $X_j^e(\tau)$ se interpreta como una función escalonada por partes, constante en intervalos unitarios correspondientes a cada año relativo. Esta equivalencia permite que

el enfoque funcional se mantenga válido y coherente, incluso cuando los datos disponibles son observaciones anuales discretas.

Este criterio es ampliamente utilizado en aplicaciones de Análisis Funcional de Datos, donde funciones reales son representadas mediante trayectorias discretas evaluadas sobre rejillas temporales finitas. En este caso, cada indicador financiero puede verse como una curva escalonada en $\mathbb{R}^3$, cuya evolución se define únicamente en puntos específicos de tiempo relativo. En la siguiente figura se ilustra un ejemplo tridimensional de esta representación escalonada.

3.3.1. Ejemplo ilustrativo del cálculo de la métrica funcional

Para facilitar la comprensión de las etapas que componen la métrica funcional propuesta, se presenta a continuación un ejemplo didáctico utilizando trayectorias sintéticas. Este ejemplo tiene fines exclusivamente ilustrativos y no representa datos reales.

Se consideran dos empresas ficticias, e_1 y e_2. Las matrices discretizadas correspondientes a sus trayectorias financieras, organizadas en una ventana de cinco años hacia atrás, son las siguientes:

$$\mathbf{x}^{e_1} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ \infty & \infty & \infty & 4 & 10 \\ 4 & 5 & 0 & 1 & 1 \end{bmatrix} \qquad \mathbf{x}^{e_2} = \begin{bmatrix} \text{NaN} & 4 & 5 & 6 & 3 \\ \text{NaN} & 2 & 1 & 2 & 6 \\ \text{NaN} & 3 & 4 & 4 & 2 \end{bmatrix}$$

donde cada fila representa un indicador financiero (ROA, LC y CTN, en ese orden), y cada columna corresponde a un instante de tiempo relativo $t = -4, \ldots, 0$.
En este ejemplo, la empresa e_2 solo dispone de información para los últimos cuatro años relativos (de $t = -3$ a $t = 0$), lo que se refleja en los valores faltantes (`NaN`) en el primer año de cada indicador. Por su parte, la empresa e_1 presenta información completa para toda la ventana, aunque en el caso de la Liquidez Corriente (segunda fila), los tres primeros valores son infinitos, lo que representa situaciones contables reales sin necesidad de imputación. Estos valores extremos son tratados mediante una transformación acotada, sin ser descartados ni modificados.

Los parámetros utilizados en este ejemplo son:

- Longitud de la ventana: $\ell = 5$
- Penalización por pérdida de dominio: $\lambda = 1$
- Pesos por indicador: $w_1 = w_2 = w_3 = \frac{1}{3}$

Paso 1: Cálculo de distancias acumuladas por indicador. Dado que la empresa e_2 solo tiene datos disponibles desde $t = -3$, el subintervalo común entre ambas empresas es $t \in [-3, 0]$. En este intervalo, se calcula la distancia acumulada tipo L^1 para cada indicador como la suma de diferencias absolutas año a año:

$$d_1 = \sum_{t=-3}^{0} \left| x_{1,t}^{e_1} - x_{1,t}^{e_2} \right| = 8$$

$$d_2 = \sum_{t=-3}^{0} \left| x_{2,t}^{e_1} - x_{2,t}^{e_2} \right| = \infty$$

$$d_3 = \sum_{t=-3}^{0} \left| x_{3,t}^{e_1} - x_{3,t}^{e_2} \right| = 10$$

donde los valores infinitos en el segundo indicador (LC) de e_1 generan una distancia infinita en ese componente.

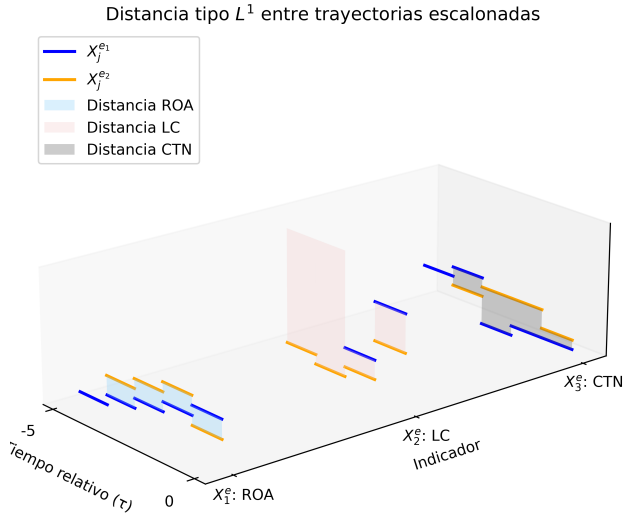

Figura 3.7: Las áreas sombreadas muestran las diferencias acumuladas por indicador entre e_1 y e_2.
Fuente: Elaboración propia.

La Figura 3.7 ilustra visualmente estas trayectorias, organizadas por indicador en un espacio tridimensional. Las áreas sombreadas representan las distancias tipo L^1 acumuladas antes de aplicar penalización o transformación.

Paso 2: Penalización por pérdida de dominio. La penalización se aplica como un factor multiplicativo cuando una trayectoria no cubre toda la ventana de observación. En este caso, la empresa e_2 pierde un año de dominio, ya que solo se tiene información desde $t = -3$. La penalización se calcula como:

$$\text{Penalización} = 1 + \lambda \cdot \frac{k_j + \ell - 1}{\ell} = 1 + 1 \cdot \frac{-3 + 5 - 1}{5} = 1 + \frac{1}{5} = 1.2$$

Aplicando esta penalización a cada una de las distancias acumuladas del paso anterior, se obtiene:

$$d_1^{(pen)} = 8 \cdot 1.2 = 9.6, \qquad d_2^{(pen)} = \infty \cdot 1.2 = \infty, \qquad d_3^{(pen)} = 10 \cdot 1.2 = 12$$

La penalización resultante incrementa proporcionalmente la distancia cuando una de las trayectorias presenta pérdida de dominio, como ocurre en este caso con e_2, que carece de información en el instante $t = -4$. Esta situación se representa gráficamente en la Figura 3.8, donde se destaca el tramo penalizado como una proporción adicional sobre la distancia original.

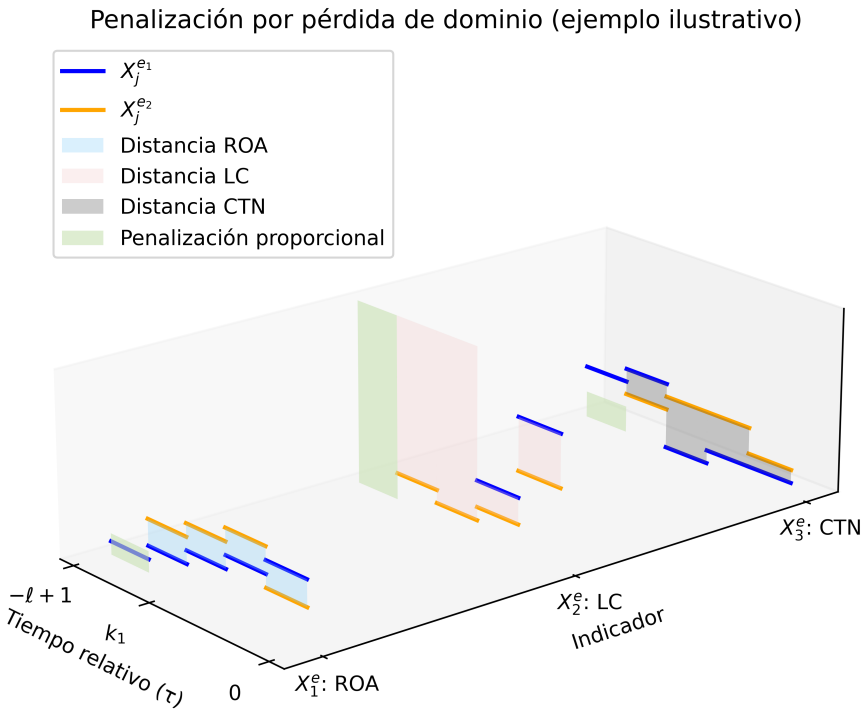

Figura 3.8: Ejemplo de penalización proporcional aplicada cuando una trayectoria tiene datos faltantes. La sección faltante se penaliza como fracción del dominio perdido.
Fuente: Elaboración propia.

Paso 3: Transformación acotada de la distancia. Para evitar que valores extremos dominen el cálculo de la distancia total, cada componente penalizado se transforma mediante la siguiente función acotadora:

$$\tilde{d}_j = \frac{d_j^{(pen)}}{1 + d_j^{(pen)}}$$

Aplicando esta transformación a los valores obtenidos en el paso anterior, se

tiene:

$$\tilde{d}_1 = \frac{9.6}{1+9.6} = \frac{9.6}{10.6} \approx 0.9057, \qquad \tilde{d}_2 = \frac{\infty}{1+\infty} = 1, \qquad \tilde{d}_3 = \frac{12}{1+12} = \frac{12}{13} \approx 0.9231$$

La aplicación de esta transformación acotadora permite que todas las distancias, incluidas aquellas con valores extremos o infinitos, queden contenidas en el intervalo $[0, 1]$, asignando $\tilde{d}_j = 1$ cuando la distancia penalizada $d_j^{(\text{pen})}$ es infinita. Esto favorece la estabilidad del modelo sin necesidad de normalizar previamente los datos financieros.

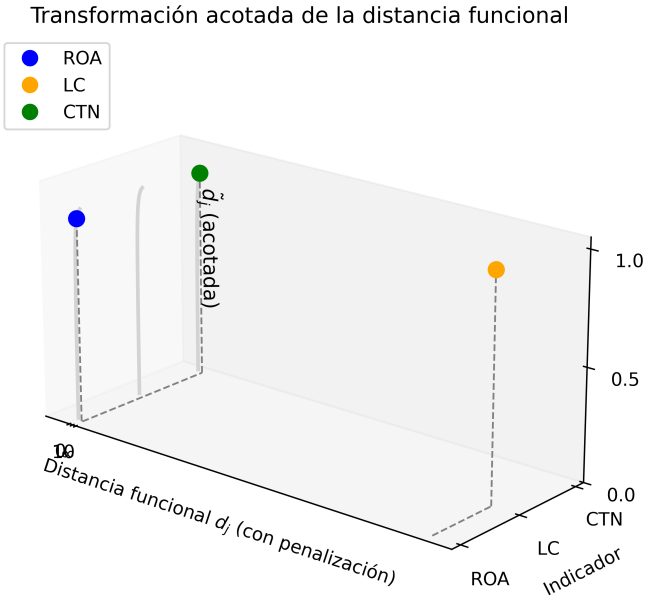

Figura 3.9: Transformación acotada de la distancia mediante $\tilde{d}_j = \frac{d}{1+d}$. Se evita que valores extremos dominen la distancia total.
Fuente: Elaboración propia.

Paso 4: Combinación ponderada de distancias. Una vez transformadas las distancias penalizadas en su versión acotada, se calcula la distancia funcional total entre las dos empresas mediante una combinación ponderada:

$$D(e_1, e_2) = \sum_{j=1}^{m} w_j \cdot \tilde{d}_j$$

donde $w_j \in [0, 1]$ representa el peso asignado al indicador j, con la restricción $\sum_{j=1}^{m} w_j = 1$. En este ejemplo, se emplean pesos iguales para cada uno de los tres indicadores, es decir, $w_1 = w_2 = w_3 = \frac{1}{3}$.

$$D(e_1, e_2) = \frac{1}{3}(0.9057 + 1 + 0.9231) = \frac{2.8288}{3} \approx 0.9429$$

Este valor representa la distancia funcional total entre las trayectorias de e_1 y e_2, considerando tanto sus diferencias acumuladas como los mecanismos de penalización y acotamiento implementados previamente.
La Figura 3.10 muestra gráficamente el resultado de esta combinación ponderada de distancias acotadas, donde cada indicador contribuye con su respectivo valor ajustado.

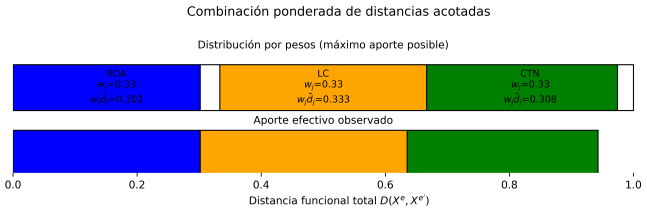

Figura 3.10: Cálculo final de la distancia total como combinación ponderada de distancias acotadas por indicador.
Fuente: Elaboración propia.

Ejemplo aplicado con datos reales. Para ilustrar la métrica funcional con trayectorias reales, se seleccionaron dos empresas del sector turismo en Colombia. La primera, con NIT **900258258.0**, presenta información financiera completa durante los últimos cinco años (2019–2023). La segunda, con NIT **805025185.0**, reporta únicamente cuatro años de datos válidos (2020–2023), reflejando una pérdida de dominio de un año. Las matrices de trayectorias discretizadas para cada empresa son

Tabla 3.1: *Trayectoria funcional de la empresa 900258258.0 (completa)*

Indicador	-4	-3	-2	-1	0
AFT	1.398	1.407	1.161	1.493	1.064
KTNO	2.65e6	2.93e6	2.02e5	-9.86e5	4.44e5
LG	2.933	4.116	5.073	3.910	2.986
MB	0.620	0.635	0.600	0.600	0.665
MO	0.132	0.112	0.164	0.151	0.189
PPI	13.15	18.51	6.39	11.64	9.685
PPP	0.000	0.000	0.000	33.77	0.000
RA	1.465	0.765	1.486	1.818	1.961
RAF	0.988	0.917	1.403	1.752	1.043
RAO	0.546	0.370	0.732	0.718	0.758
RCC	3.902	1.845	307.15	307.15	307.15
RCID	0.301	0.134	0.419	0.391	0.697
RCP	1315.0	1315.0	1315.0	10.81	1315.0
RI	27.76	19.72	57.12	31.35	37.69
ROE	0.197	0.047	0.284	0.423	0.334
ROI	0.546	0.259	0.415	0.526	0.756
T	0.311	0.311	0.641	0.574	0.768

Tabla 3.2: *Trayectoria funcional de la empresa 805025185.0 (incompleta)*

Indicador	-4	-3	-2	-1	0
AFT	NaN	1.631	1.415	1.393	1.464
KTNO	NaN	1.58e4	2.16e4	5.51e4	2.87e4
LG	NaN	1.217	1.320	1.395	1.352
MB	NaN	0.869	0.893	0.886	0.879
MO	NaN	0.132	0.023	0.031	0.021
PPI	NaN	3.600	3.354	7.290	3.776
PPP	NaN	0.000	0.000	0.000	0.000
RA	NaN	0.356	0.615	0.765	0.741
RAF	NaN	3.629	3.099	3.046	3.092
RAO	NaN	0.895	0.465	0.582	0.339
RCC	NaN	307.15	307.15	307.15	307.15
RCID	NaN	0.059	0.019	0.032	0.021
RCP	NaN	1315.0	1315.0	1315.0	1315.0
RI	NaN	101.39	108.82	50.07	96.67
ROE	NaN	0.184	-0.038	0.015	-0.044
ROI	NaN	0.064	0.017	0.028	0.020
T	NaN	0.212	0.134	0.139	0.217

Los resultados del cálculo de distancias por indicador se resumen en la Tabla 3.3, incluyendo el valor L^1, el número de faltantes, la distancia penalizada y su transformación acotada.

La distancia funcional entre ambas empresas se calculó con penalización $\lambda = 1.0$ y pesos iguales para cada indicador. Los resultados obtenidos por indicador son:

Indicador	L1	Faltantes	Penalizada	Acotada
AFT	0.9795	1	1.1754	0.5403
KTNO	4.55e+06	1	5.46e+06	1.0000
LG	10.8024	1	12.9629	0.9284
MB	1.0260	1	1.2312	0.5518
MO	0.4490	1	0.5388	0.3502
PPI	28.2072	1	33.8486	0.9713
PPP	33.7689	1	40.5226	0.9759
RA	3.5537	1	4.2644	0.8100
RAF	7.7503	1	9.3004	0.9029
RAO	1.3469	1	1.6163	0.6178
RCC	305.3034	1	366.3641	0.9973
RCID	1.5089	1	1.8106	0.6442
RCP	1304.1910	1	1565.0290	0.9994
RI	211.0712	1	253.2855	0.9961
ROE	1.2451	1	1.4941	0.5990
ROI	1.8270	1	2.1924	0.6868
T	1.5918	1	1.9102	0.6564

Tabla 3.3: *Cálculo de la distancia funcional por indicador entre las empresas 900258258.0 y 805025185.0.*

La distancia funcional total obtenida fue:

$$D(900258258.0, 805025185.0) = 0.7781$$

Este mecanismo garantiza una comparación equitativa entre empresas con distinta longitud de historial financiero, sin recurrir a imputaciones o descartes agresivos de datos.

3.4. Extensiones posibles de la métrica funcional personalizada

La métrica funcional personalizada desarrollada en esta investigación fue concebida para comparar trayectorias financieras multivariadas de longitud homogénea, capturando similitudes temporales relevantes entre empresas. Sin embargo, su formulación original presenta una limitación importante: no incorpora variables categóricas o atributos estructurales constantes, los cuales han demostrado ser determinantes para mejorar la capacidad predictiva en modelos de quiebra empresarial.

Diversos estudios recientes respaldan esta afirmación. Por ejemplo, Gnip et al. (2025) y Dasilas y Rigani (2024) muestran que la inclusión de características cualitativas como el sector económico, la jurisdicción o el tipo societario incrementa significativamente el rendimiento de modelos como XGBoost o LightGBM. De forma complementaria, la revisión de Kuizinienė et al. (2022) destaca que más del 60 % de los modelos más efectivos en los últimos años combinan datos contables con variables categóricas, incluso en contextos con alta heterogeneidad o desbalanceo de clases.

En este sentido, una posible línea de extensión consiste en adaptar la métrica propuesta para que pueda incorporar este tipo de información estructural. Esto puede lograrse mediante codificaciones explícitas —como variables *dummy*— o mediante el diseño de funciones de distancia específicas para componentes categóricos. Esta ampliación enriquecería la representación del espacio funcional, permitiendo capturar dimensiones adicionales del entorno empresarial que inciden en el comportamiento financiero de las firmas.

Finalmente, estas evidencias respaldan la idea de que una métrica de similitud entre empresas no debe depender exclusivamente de sus trayectorias contables, sino que puede beneficiarse de integrar componentes estructurales y contextuales. De hecho, diversos trabajos recientes han propuesto marcos compuestos o híbridos que combinan distancias funcionales con atributos cualitativos o macroeconómicos, obteniendo así métricas más robustas y representativas del entorno real (Kar et al., 2024; P. Liu et al., 2024; Thompson & Davison, 2024).

3.4.1. Extensión 1: Incorporación de variables categóricas estáticas

La métrica funcional propuesta en esta investigación fue diseñada originalmente sobre un espacio funcional $\mathcal{F}$, compuesto por trayectorias multivariadas de indicadores financieros. Sin embargo, en muchas aplicaciones prácticas es relevante considerar atributos estructurales que no evolucionan en el tiempo, pero que pueden incidir de manera significativa en el comportamiento de las empresas. Ejemplos de estas variables incluyen la clasificación sectorial (CIIU), el departamento geográfico (DEP), o el tipo societario. Estas variables suelen ser categóricas y se mantienen constantes durante el periodo de observación.

Para integrar este tipo de información, se propone ampliar el espacio de representación de las empresas mediante un producto cartesiano entre el espacio funcional $\mathcal{F}$ y un nuevo espacio categórico $\mathcal{S}$. El espacio total se denota como:

$$\mathcal{X} = \mathcal{S} \times \mathcal{F}$$

De esta forma, cada empresa e es representada mediante un par ordenado compuesto por dos componentes:

$$\mathbf{x}^e = (\mathbf{x}^e_{\mathcal{S}},\, \mathbf{x}^e_{\mathcal{F}})$$

- $\mathbf{x}^e_{\mathcal{F}} \in \mathbb{R}^{m\times\ell}$ corresponde a la matriz de trayectorias de m indicadores financieros observados durante ℓ periodos.
- $\mathbf{x}^e_{\mathcal{S}} \in \mathcal{C}^p$ representa un vector de p atributos categóricos estáticos, donde $\mathcal{C}$ es un conjunto de valores no numéricos.

Por ejemplo, si se consideran dos variables categóricas estáticas (por ejemplo, sector económico y ubicación geográfica), su representación general sería:

$$\mathbf{x}^e_{\mathcal{S}} = \begin{bmatrix} x^e_{\text{cat}_1} \\ x^e_{\text{cat}_2} \end{bmatrix} \quad \text{y} \quad \mathbf{x}^e_{\mathcal{F}} = \begin{bmatrix} x^e_{1,-\ell+1} & x^e_{1,-\ell+2} & \cdots & x^e_{1,0} \\ x^e_{2,-\ell+1} & x^e_{2,-\ell+2} & \cdots & x^e_{2,0} \\ \vdots & \vdots & & \vdots \\ x^e_{m,-\ell+1} & x^e_{m,-\ell+2} & \cdots & x^e_{m,0} \end{bmatrix}$$

Esta formulación permite mantener separados los componentes funcionales y categóricos, lo cual facilita el uso de funciones de distancia especializadas

para cada tipo de variable. En particular, las variables en $\mathcal{S}$ deben ser tratadas mediante funciones de similitud o distancia acordes con su naturaleza categórica o discreta, de forma que reflejen correctamente el grado de coincidencia o divergencia entre empresas en dichos atributos. Por su parte, la parte funcional continúa utilizando la métrica personalizada definida en esta tesis. De esta manera, se conserva la trazabilidad del modelo y se amplía su aplicabilidad a contextos más realistas.

3.4.2. Extensión 2: Incorporación de variables categóricas dinámicas

En algunas aplicaciones, existen variables categóricas que no permanecen constantes a lo largo del tiempo, sino que evolucionan durante la trayectoria de la empresa. Este es el caso, por ejemplo, del régimen tributario, el tipo societario o la calificación crediticia, los cuales pueden cambiar entre periodos y reflejar transiciones estructurales relevantes. A diferencia de las variables estáticas del espacio $\mathcal{S}$, estas variables categóricas presentan una dinámica temporal que debe ser representada mediante trayectorias discretas.

Para ello, se propone ampliar nuevamente el espacio de representación funcional mediante un producto cartesiano entre el espacio $\mathcal{F}$ y un nuevo espacio categórico dinámico $\mathcal{S}_{\text{cat}}$, que mantiene la estructura matricial pero cuyas entradas no son valores reales sino categorías. El espacio total se denota:

$$\mathcal{X} = \mathcal{S}_{\text{cat}} \times \mathcal{F}$$

Cada empresa e es representada mediante un par ordenado con dos componentes:

$$\mathbf{x}^e = \left(\mathbf{x}^e_{\mathcal{S}_{\text{cat}}}, \mathbf{x}^e_{\mathcal{F}}\right)$$

- $\mathbf{x}^e_{\mathcal{F}} \in \mathbb{R}^{m \times \ell}$ representa la trayectoria multivariada de indicadores financieros.
- $\mathbf{x}^e_{\mathcal{S}_{\text{cat}}} \in \mathcal{C}^{q \times \ell}$ representa un conjunto de q trayectorias categóricas dinámicas, observadas durante ℓ periodos.

Por ejemplo, si se consideran dos variables categóricas dinámicas, su representación general sería:

$$\mathbf{x}^e_{\mathcal{S}_{\text{cat}}} = \begin{bmatrix} c^e_{1,-\ell+1} & c^e_{1,-\ell+2} & \cdots & c^e_{1,0} \\ c^e_{2,-\ell+1} & c^e_{2,-\ell+2} & \cdots & c^e_{2,0} \end{bmatrix} \quad \text{y} \quad \mathbf{x}^e_{\mathcal{F}} = \begin{bmatrix} x^e_{1,-\ell+1} & x^e_{1,-\ell+2} & \cdots & x^e_{1,0} \\ x^e_{2,-\ell+1} & x^e_{2,-\ell+2} & \cdots & x^e_{2,0} \\ \vdots & \vdots & & \vdots \\ x^e_{m,-\ell+1} & x^e_{m,-\ell+2} & \cdots & x^e_{m,0} \end{bmatrix}$$

Esta formulación mantiene la lógica de separación entre componentes funcionales y categóricos, permitiendo aplicar funciones de distancia distintas para cada tipo. En particular, las trayectorias categóricas pueden ser evaluadas mediante medidas de coincidencia, secuencias de cambio o distancias adaptadas a datos nominales, sin necesidad de transformarlas a valores continuos. Al combinar esta información con la métrica funcional, se obtiene una representación más completa y contextual del comportamiento empresarial.

3.4.3. Extensión 3: Incorporación de variables cuantitativas estáticas

Además de los atributos categóricos, existen variables numéricas que caracterizan a la empresa de forma estructural pero que no cambian a lo largo de la ventana de observación. Estas variables cuantitativas estáticas pueden incluir, por ejemplo, la antigüedad de la empresa, el número total de empleados, el nivel de activos fijos o el promedio histórico de ingresos. Su inclusión puede ser útil para capturar diferencias estructurales de magnitud o escala que influyen en el riesgo de quiebra, pero que no forman parte de la trayectoria temporal.

Para representarlas, se introduce un nuevo espacio $\mathcal{Q} \subset \mathbb{R}^r$, donde r corresponde al número de atributos cuantitativos estáticos. El espacio total se amplía nuevamente como:

$$\mathcal{X} = \mathcal{Q} \times \mathcal{F}$$

Cada empresa e es entonces representada mediante un par ordenado:

$$\mathbf{x}^e = \left(\mathbf{x}^e_{\mathcal{Q}},\, \mathbf{x}^e_{\mathcal{F}}\right)$$

- $\mathbf{x}^e_{\mathcal{F}} \in \mathbb{R}^{m \times \ell}$ representa la trayectoria funcional multivariada.
- $\mathbf{x}^e_{\mathcal{Q}} \in \mathbb{R}^r$ corresponde a un vector de valores numéricos fijos por empresa.

Por ejemplo, si se consideran r variables cuantitativas estáticas, su representación general sería:

$$\mathbf{x}^e_{\mathcal{Q}} = \begin{bmatrix} x^e_{\mathrm{q}_1} \\ \vdots \\ x^e_{\mathrm{q}_r} \end{bmatrix} \quad \text{y} \quad \mathbf{x}^e_{\mathcal{F}} = \begin{bmatrix} x^e_{1,-\ell+1} & x^e_{1,-\ell+2} & \cdots & x^e_{1,0} \\ x^e_{2,-\ell+1} & x^e_{2,-\ell+2} & \cdots & x^e_{2,0} \\ \vdots & \vdots & & \vdots \\ x^e_{m,-\ell+1} & x^e_{m,-\ell+2} & \cdots & x^e_{m,0} \end{bmatrix}$$

La inclusión de estas variables en el modelo permite combinar la evolución temporal con condiciones estructurales cuantificables, enriqueciendo así la capacidad explicativa del análisis. En este caso, las variables en $\mathcal{Q}$ pueden ser tratadas mediante distancias euclidianas o métricas normalizadas, y luego integradas con la distancia funcional a través de un esquema de combinación ponderada como el planteado en esta tesis.

Las tres extensiones descritas —incorporación de variables categóricas estáticas, categóricas dinámicas y cuantitativas estáticas— representan vías complementarias para enriquecer el espacio funcional $\mathcal{F}$ mediante la integración de atributos estructurales y contextuales. En escenarios reales, es común que coexistan múltiples tipos de variables relevantes: por ejemplo, una empresa puede estar ubicada en un determinado departamento (*categórica estática*), pertenecer a un segmento de mercado que cambia con el tiempo (*categórica dinámica*) y contar con un número fijo de empleados o años de antigüedad al momento de la observación (*cuantitativa estática*). En estos casos, el espacio total de representación puede formularse como una combinación de subespacios heterogéneos:

$$\mathcal{X} = \mathcal{S} \times \mathcal{T} \times \mathcal{Q} \times \mathcal{F}$$

donde $\mathcal{S}$, $\mathcal{T}$ y $\mathcal{Q}$ corresponden respectivamente a espacios categóricos estáticos, categóricos dinámicos y cuantitativos estáticos. Cada empresa queda representada como un cuádruple ordenado:

$$\mathbf{x}^e = \left(\mathbf{x}^e_{\mathcal{S}}, \mathbf{x}^e_{\mathcal{T}}, \mathbf{x}^e_{\mathcal{Q}}, \mathbf{x}^e_{\mathcal{F}}\right)$$

Esta generalización ofrece un marco flexible para construir métricas compuestas, en las que cada componente puede ser tratado con una función de distancia acorde con su naturaleza. Por ejemplo, las variables categóricas pueden evaluarse mediante distancias tipo coincidencia simple, similitud de Jaccard o coeficientes como Gower, mientras que las cuantitativas estáticas

pueden compararse usando métricas euclidianas o de Mahalanobis tras una normalización adecuada. La clave metodológica consiste en estudiar cuidadosamente el significado, escala y comportamiento esperado de cada tipo de variable para seleccionar (o diseñar) una función de similitud que preserve su capacidad explicativa sin introducir sesgos indebidos.

La métrica funcional personalizada propuesta en esta tesis puede, por tanto, ampliarse hacia un enfoque mixto con alto potencial explicativo y adaptable a múltiples dominios. Esta dirección de trabajo abre oportunidades para construir modelos más realistas, sensibles al contexto estructural de las empresas, y capaces de capturar patrones complejos en entornos empresariales heterogéneos.

Capítulo 4

Procesamiento de datos, imputación contable y generación de variables predictoras

Este capítulo describe el proceso completo de preparación de los datos utilizados en esta investigación, desde la recolección de información financiera hasta la construcción de las variables necesarias para el análisis funcional y la predicción de quiebra empresarial. El objetivo principal es garantizar la coherencia, integridad y relevancia de los datos que alimentarán el modelo funcional propuesto, respetando tanto criterios contables como consideraciones metodológicas.

Se parte de una base amplia de estados financieros de empresas colombianas, reportados oficialmente ante la Superintendencia de Sociedades, y se avanza hacia la depuración, imputación y transformación de dichos datos en trayectorias multivariadas que reflejan la evolución temporal de los indicadores clave.

Aunque el enfoque del análisis se centra en el sector turismo, las etapas iniciales del procesamiento —incluyendo la consolidación, estandarización y limpieza de los datos— se aplicaron sobre el universo completo de empresas. Esta estrategia busca asegurar la replicabilidad del flujo de trabajo en futuros estudios sectoriales, evitando la pérdida de información relevante durante los pasos críticos de verificación e imputación. El filtrado por sector se realiza posteriormente, como una etapa específica y controlada, una vez garantizada

la consistencia general de la base de datos.

4.1. Fuentes de información y cobertura temporal

La fuente de información utilizada en este estudio corresponde a los estados financieros anuales reportados por las empresas colombianas ante la Superintendencia de Sociedades. Los datos fueron descargados directamente del Sistema Integrado de Información Societaria (SIIS)[1], una plataforma oficial que permite la consulta y descarga masiva de archivos financieros consolidados por año.

El conjunto de datos abarca el periodo 1995–2023, los estados financieros descargados presentan tres estructuras distintas según el año. En primer lugar, para el periodo comprendido entre 1995 y 2003, así como el año 2007, los archivos están consolidados por año, generalmente en uno o dos libros de Excel, e incluyen tanto el balance general como el estado de resultados y el flujo de efectivo. En segundo lugar, para los años entre 2004 y 2015 (excluyendo 2007), los datos se presentan de forma separada en tres archivos por año, correspondientes al Balance General, Estado de Resultados y Flujo de Efectivo, siguiendo la nomenclatura estandarizada del sistema SIREM. Finalmente, desde 2016 hasta 2023, los archivos están clasificados por tipo de empresa (Plenas y PYMES), y cada año contiene seis archivos diferenciados por información: carátula, situación financiera, estado de resultados, resultado integral, flujo de efectivo y cambios en el patrimonio. Esta última estructura responde a los lineamientos de las Normas Internacionales de Información Financiera (NIIF), reflejando una mayor desagregación y estandarización.

El uso exclusivo de esta fuente se justifica plenamente por su carácter oficial, obligatorio y estandarizado. La Superintendencia de Sociedades es el organismo competente para supervisar la información financiera de las empresas colombianas que superan umbrales específicos de activos o ingresos. Por tanto, los datos del SIIS constituyen el universo de referencia para los análisis financieros y regulatorios en el país, siendo también la base utilizada en estudios académicos recientes, como en Correa (2023), y construida con base en registros administrativos reportados a la Superintendencia de Sociedades de Colombia.

Aunque la adopción de las Normas Internacionales de Información Financiera (NIIF) a partir de 2016 implicó cambios sustanciales en la forma de

[1]Superintendencia de Sociedades. (s.f.). *Sistema Integrado de Información Societaria (SIIS)*. Recuperado el 3 de mayo de 2025, de https://siis.ia.supersociedades.gov.co/#/

presentar los estados financieros en Colombia, este estudio no excluyó los datos previos. A pesar de las diferencias formales entre periodos, las cuentas contables clave —como activos, pasivos, ingresos, gastos y patrimonio— conservan una continuidad conceptual que permite su estandarización técnica mediante un proceso riguroso de limpieza y normalización.

4.1.1. Imputación contable estructurada de cuentas financieras

Antes del análisis, se ejecutó un proceso riguroso de limpieza y validación estructural de los datos financieros extraídos del SIIS. Este proceso incluyó la verificación de la legibilidad de los archivos por año, la existencia de las principales tablas contables (balance general, estado de resultados y flujo de efectivo), y la coherencia interna entre ellas.

Se desarrolló un procedimiento automatizado en Python para consolidar los archivos descargados en una base unificada, permitiendo integrar información proveniente de múltiples formatos y regímenes contables. En particular, se tomaron en cuenta las diferencias entre los marcos contables previos y posteriores a la adopción de las NIIF a partir de 2016.

Durante la estandarización, se resolvieron problemas comunes como nombres de columnas inconsistentes, registros incompletos y codificaciones sectoriales heterogéneas. Se normalizaron variables clave como el identificador de empresa (NIT), el año, la ciudad, el departamento y el código CIIU. Además, se aplicaron reglas contables para imputar valores faltantes de forma controlada, garantizando la coherencia entre activos, pasivos y patrimonio.

Adicionalmente, se verificó la consistencia del identificador empresarial (NIT) entre los distintos estados financieros de cada año, identificando y corrigiendo discrepancias en años específicos (como 2001). Esta revisión garantizó que los datos consolidados por empresa fueran coherentes entre balance general, estado de resultados y flujo de efectivo. También se eliminaron registros duplicados por empresa y año, conservando únicamente la primera ocurrencia válida, con el fin de asegurar una única observación por unidad de análisis.

En una etapa posterior, se llevó a cabo un proceso sistemático de limpieza geográfica. Inicialmente se identificaron 25.597 registros sin departamento reportado. A través de la estandarización de nombres de ciudad y la imputación por correspondencia, se logró reducir esta cifra a 2.199 casos. Posteriormente, se completaron los valores restantes mediante la asignación del departamento más frecuente por NIT, lo que dejó un total de tan solo 11 registros sin información departamental. En paralelo, se limpiaron y estandarizaron 566 nombres únicos de ciudad, eliminando inconsistencias ortográficas, espacios redundantes y errores tipográficos.

En cuanto a la clasificación sectorial, se implementó un esquema de codificación a partir del código CIIU, extrayendo la letra y los dos primeros

dígitos para asignar cada empresa a uno de los 21 sectores económicos definidos por la normativa nacional. Este procedimiento se aplicó con éxito a la totalidad de los registros, sin que quedaran observaciones sin clasificación sectorial al finalizar la limpieza.

Finalmente, se construyó la variable binaria `Turismo`, que identifica si una empresa pertenece o no al sector turístico. Esta clasificación se realizó con base en listas específicas de códigos CIIU pertenecientes a las revisiones 3 y 4, según el año del reporte. Como resultado, se identificaron 5.839 empresas turísticas (NIT únicos), con un total de 52.720 registros anuales relacionados.

Tras eliminar 14 registros con datos esenciales faltantes (como departamento o clasificación sectorial), la base final quedó compuesta por 542.069 registros válidos, lo cual garantiza una estructura depurada, coherente y lista para el análisis posterior de trayectorias empresariales y riesgo de quiebra.

4.1.2. Imputación técnica de cuentas financieras

Con el fin de garantizar la completitud de las variables necesarias para el cálculo de indicadores financieros, se implementó un procedimiento de imputación de datos faltantes sobre las principales cuentas contables. La imputación se realizó con base en criterios contables lógicos, consistencia interna entre variables y evidencia histórica por empresa.

En una primera etapa, se imputaron con valor cero aquellas cuentas cuyo valor faltante puede interpretarse razonablemente como ausencia de movimiento, como es el caso de inventarios, cuentas por cobrar, cuentas por pagar, gastos financieros, gastos de ventas, gastos administrativos, depreciaciones y amortizaciones, y flujo de caja libre. Esta imputación preserva la coherencia de las ecuaciones contables, ya que el valor cero no introduce distorsiones entre ingresos, costos, utilidades o flujos.

Posteriormente, otras cuentas como activos fijos, impuestos, utilidad neta, EBIT y utilidad bruta fueron imputadas mediante reglas contables verificables. Por ejemplo, se imputó la utilidad neta como la diferencia entre EBIT e impuestos cuando ambos valores estaban disponibles, o como cero si no existían ingresos. También se reconstruyeron variables como la utilidad bruta o la utilidad operacional a partir de componentes disponibles, como ventas, gastos administrativos y financieros.

En una segunda etapa, se imputaron con ceros aquellos valores aún faltantes que, aunque no cumplían directamente una regla contable, habían sido reportados por la misma empresa (NIT) en otros años. Esta imputación

basada en evidencia histórica permitió recuperar un número importante de observaciones sin recurrir a métodos automáticos.

No se aplicaron técnicas de imputación estadística como MICE, k-Nearest Neighbors o regresiones múltiples, ya que estas pueden generar combinaciones inconsistentes desde el punto de vista contable, alterando relaciones fundamentales entre activos, pasivos, patrimonio, ingresos y utilidades. En cambio, se priorizó la imputación basada en lógica financiera, reglas explícitas y validación interna.

De manera complementaria, se presenta un resumen cuantitativo del proceso. Antes de la depuración, variables como Depreciaciones y Amortizaciones, Costos de Producción o Flujo de Caja Libre presentaban más de 100.000 valores faltantes. Tras aplicar las reglas contables, la imputación con ceros y la validación por historial, el número de registros con al menos un valor perdido se redujo a solo 2.751, es decir, el 0,51 % del total. Estas observaciones residuales fueron eliminadas para garantizar una base completamente depurada y coherente.

Como resultado, se obtuvo una base final compuesta por 539.318 registros válidos, sin valores nulos en ninguna de las 22 cuentas contables principales. Esta condición garantiza una estructura sólida para el cálculo posterior de indicadores financieros y la implementación de modelos predictivos.

4.2. Cálculo, imputación y estructuración de indicadores financieros

Sobre la base imputada y normalizada en la etapa anterior, se procedió al cálculo de 45 indicadores financieros estandarizados que capturan distintos aspectos de la liquidez, endeudamiento, rentabilidad, eficiencia y productividad empresarial. Estos indicadores fueron construidos a partir de relaciones contables ampliamente reconocidas en la literatura financiera y adaptadas a la estructura de datos disponible.

Cada indicador fue calculado considerando condiciones especiales para evitar distorsiones numéricas y preservar la lógica económica subyacente. Por ejemplo, para la razón de liquidez corriente (LC), si tanto los activos corrientes como los pasivos corrientes eran cero, el valor se consideró indefinido (`NaN`). En los casos en que los pasivos eran cero pero existían activos, se asignó un valor infinito positivo (∞), reflejando una empresa completamente líquida. Cuando los pasivos eran positivos, se aplicó la fórmula directa: activos sobre pasivos corrientes. Esta lógica se extendió a otros indicadores como la prueba ácida (PA), el capital de trabajo neto (CTN) o la razón de apalancamiento financiero (RAF). En todos los casos se registró la distribución de valores normales, indefinidos, infinitos o negativos, según su pertinencia financiera.

A continuación, se aplicó una imputación diferenciada de los valores faltantes con fines predictivos. En lugar de usar métodos automáticos, se adoptó una estrategia basada en la interpretación contable de cada indicador. Se asignó el valor 1 a aquellos indicadores donde representa un equilibrio o punto neutro (como en LC, PA o T), el valor 0 cuando un resultado nulo refleja ausencia de rentabilidad o eficiencia, y el valor 0.5 en algunos casos donde no se puede determinar una proporción dominante entre deuda y patrimonio.

En cuanto al tratamiento de valores extremos, los valores infinitos positivos se conservaron siempre que fueran coherentes con la lógica económica del indicador. Por ejemplo, un valor de LC igual a ∞ puede reflejar una empresa sin pasivos corrientes pero con activos líquidos. Sin embargo, aquellos casos en los que un valor infinito implicaba relaciones contables inconsistentes —como deuda sin activos o capital sin respaldo financiero— fueron eliminados o tratados de manera específica.

Si bien el cálculo de indicadores se realizó sobre la base completa de empresas, el análisis posterior y la construcción de modelos se enfocaron exclusivamente en el sector turismo. Por ello, en la Tabla 4.1 se presenta

el resumen final de calidad de datos por tipo de valor únicamente para las observaciones correspondientes a empresas turísticas.

Tabla 4.1: *Indicadores financieros en la base del sector turismo: conteo de valores finitos y* ∞

Código	**Indicador**	$+\infty$	**Valores finitos**
AFT	Activos Fijos sobre Total de Capital	1	52,574
CCNAT	Capital de Trabajo Neto sobre Activos Totales	0	52,575
CCNCT	Capital de Trabajo Neto sobre Capital Total	0	52,575
CCT	Capital Contable sobre Pasivos Totales	0	52,575
CPA	Capital Propio sobre Activos	0	52,575
CTN	Capital de Trabajo Neto	0	52,575
CTNAT	CTN sobre Activos Totales	0	52,575
CTNCT	CTN sobre Capital Total	0	52,575
DATA	Deuda Total sobre Activos Totales	0	52,575
DCA	Deuda Corriente sobre Activos	0	52,575
DLPC	Deuda de Largo Plazo sobre Capital	0	52,575
EBITDA	Beneficio antes de Intereses, Impuestos, Depreciaciones y Amortizaciones	0	52,575
EVA	Valor Económico Agregado	0	52,575
KTNO	Capital de Trabajo Neto Operativo	0	52,575
LC	Liquidez Corriente	227	52,348
LG	Liquidez General	246	52,329
MB	Margen Bruto	144	52,431
ME	Margen EBITDA	2,360	50,215
MO	Margen Operacional	2,215	50,360
PA	Prueba Ácida	226	52,349
PCD	Palanca de Crecimiento	0	52,575

Código	Indicador	$+\infty$	Valores finitos
PKT	Productividad del Capital de Trabajo	2,925	49,650
PPC	Período Promedio de Cobro	2,008	50,567
PPI	Período Promedio de Inventarios	292	52,283
PPP	Período Promedio de Pago	1,961	50,614
RA	Rotación de Activos	0	52,575
RAF	Apalancamiento Financiero	1	52,574
RAO	Rentabilidad de la Actividad Operativa	156	52,419
RCC	Rotación de Cuentas por Cobrar	11,260	41,315
RCCP	Cobertura Pasivos Corrientes con FCL	82	52,493
RCD	Cobertura Deuda Total	26	52,549
RCDT	Cobertura Deuda Total con FCL	26	52,549
RCID	Cobertura Intereses con Deuda	42	52,533
RCLP	Cobertura Pasivos No Corrientes con FCL	15,987	36,588
RCP	Rotación de Cuentas por Pagar	9,733	42,842
RECP	Endeudamiento Corriente con Patrimonio	2	52,573
RELP	Endeudamiento de Largo Plazo	0	52,575
RET	Razón de Endeudamiento Total	2	52,573
RI	Rotación de Inventarios	19,214	33,361
ROA	Rentabilidad sobre Activos Totales	0	52,575
ROE	Rentabilidad sobre Patrimonio	2	52,573
ROI	Retorno sobre la Inversión Total	2,686	49,889
RSV	Rentabilidad sobre Ventas	2,464	50,111
T	Tesorería	63	52,512
TN	Tesorería Neta	63	52,512

Nota. Se reportan los valores positivos infinitos resultantes de divisiones con denominador igual a cero, que fueron conservados cuando reflejan una situación económica coherente. La columna "Valores finitosïncluye únicamente observaciones numéricas distintas de ∞.
Fuente: elaboración propia.

4.3. Construcción de la variable dependiente (riesgo de quiebra)

En esta investigación, se construyó una variable dependiente denominada riesgo de quiebra, la cual identifica si una empresa presentó señales claras de dificultad financiera severa o descontinuidad operativa en un año determinado.

La variable fue construida a partir de dos condiciones complementarias. En primer lugar, se consideró en riesgo toda empresa que, habiendo reportado información financiera en un año determinado, no registró datos en el año siguiente, con excepción del año 2023, que corresponde al punto de corte temporal de la serie. En segundo lugar, se utilizó la información oficial de la Superintendencia de Sociedades entre 2016 y 2023, identificando como empresas en riesgo aquellas cuya condición legal no correspondía a *activa* o *en etapa preoperativa*. Estados como *acuerdo de reorganización*, *acuerdo de reestructuración*, *concordato en ejecución* o *concordato en trámite* fueron considerados evidencia directa de insolvencia, intervención o cesación de operaciones.

Esta variable se codificó en la base como una columna binaria denominada `RQ`, donde 1 indica empresas en riesgo de quiebra y 0 indica continuidad operativa.

La construcción de esta variable se encuentra en línea con enfoques recientes en la literatura académica, donde se han desarrollado definiciones operativas de quiebra o dificultad financiera basadas en registros administrativos, procesos legales o persistencia en condiciones adversas.

La construcción de esta variable encuentra respaldo en diversas investigaciones que han utilizado criterios similares para identificar empresas en riesgo financiero. En el caso colombiano, Correa (2023) propone una clasificación de quiebra para las pequeñas y medianas empresas con base en registros administrativos de la Superintendencia de Sociedades, combinando información financiera y no financiera, y demostrando que dicha etiqueta permite una predicción precisa del comportamiento empresarial. De forma complementaria, Bargagli-Stoffi et al. (2023) introducen el concepto de empresas zombis como aquellas que presentan una persistencia prolongada en condiciones de estrés financiero severo, con alta probabilidad de quiebra durante al menos tres años consecutivos; aunque esta definición se deriva de modelos predictivos y no de registros oficiales, el criterio operativo resulta análogo al empleado en este estudio. Por su parte, X. Wang y Brorsson (2024) utilizan información legal

reportada por las propias empresas —particularmente, registros de comportamiento de reestructuración— como evidencia directa de situaciones de quiebra o riesgo financiero grave, lo cual refuerza el uso de fuentes institucionales como insumos para construir este tipo de variables. Finalmente, Chi y Shen (2022) desarrollan modelos de predicción sobre empresas con dudas de continuidad operativa (*going concern doubts*), con base en variables financieras y criterios de auditoría profesional; aunque este enfoque se centra en juicios contables, el trasfondo metodológico —la evaluación de viabilidad empresarial futura— coincide con el propósito de la variable de riesgo de quiebra aquí definida.

En conjunto, estos antecedentes refuerzan la validez de la variable *riesgo de quiebra* empleada en este trabajo, al demostrar que criterios similares han sido utilizados y validados en diversos contextos internacionales y metodológicos.

Una vez construida la variable RQ, se aplicó sobre el conjunto completo de datos empresariales, antes del filtrado por sector turismo, con el fin de capturar la estructura global del fenómeno. En total, la base consolidada contiene 542.083 registros empresa-año correspondientes a 69.423 empresas únicas (NITs distintos). De estos, 71.973 registros (13,3 %) fueron clasificados como empresas en riesgo de quiebra (RQ = 1), mientras que los restantes 470.110 registros (86,7 %) corresponden a empresas operativas (RQ = 0). A nivel de empresa, se identificaron 51.275 que presentaron al menos un año con riesgo de quiebra durante el periodo observado, mientras que 18.148 no registraron nunca señales de riesgo.

Estos resultados reflejan una prevalencia significativa del riesgo financiero dentro del universo empresarial colombiano, y confirman que la variable RQ presenta una distribución empírica desbalanceada, aspecto que será considerado en las estrategias de modelamiento predictivo posteriores.

4.4. Análisis exploratorio multivariado: colinealidad y estructura latente

Antes de proceder a la construcción de modelos predictivos, se realizó un análisis exploratorio multivariado orientado a identificar y reducir la colinealidad entre los indicadores financieros. Este paso es fundamental para asegurar la validez de los supuestos en modelos paramétricos como la regresión logística, así como para mejorar la interpretabilidad y eficiencia computacional en el resto de los modelos, evitando las omisiones metodológicas aún frecuentes en la literatura reciente (Cheraghali & Molnár, 2024).

Tal como proponen Kutner et al. (2005), el cálculo del Factor de Inflación de la Varianza (VIF) permite detectar relaciones lineales fuertes entre variables explicativas, siendo valores superiores a 10 una señal clara de colinealidad problemática. Además, de acuerdo con G. James et al. (2021), la reducción dimensional por agrupamiento y selección informada de variables es una estrategia válida para preservar la interpretabilidad y robustez en modelos predictivos complejos.

Inicialmente se calcularon las correlaciones de Pearson entre todas las variables numéricas estandarizadas, construyendo un dendrograma de agrupación jerárquica que permitió identificar bloques de indicadores con relaciones redundantes o altamente lineales. La Figura 4.1 muestra el resultado de esta agrupación preliminar.

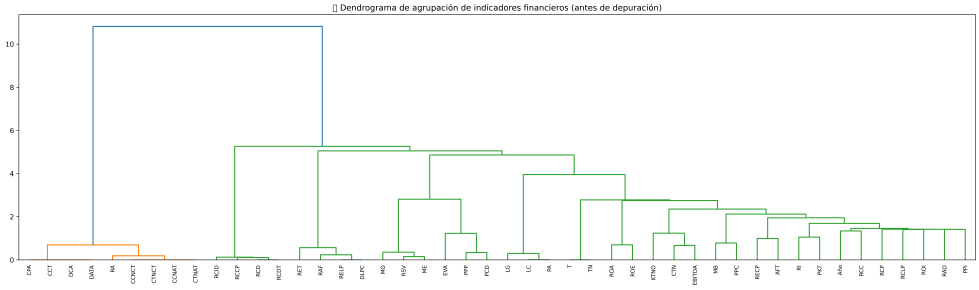

Figura 4.1: Dendrograma inicial de agrupación jerárquica entre indicadores financieros.

De forma complementaria, se calcularon los factores de inflación de la varianza (VIF) para los 45 indicadores iniciales. Se encontró que múltiples variables presentaban niveles severos de colinealidad, con VIF infinitos o

superiores a cien mil, lo que justificó la necesidad de aplicar una reducción estructurada. En la Tabla 4.2 se resumen los casos más críticos.

Tabla 4.2: *Ejemplos de variables con VIF elevado antes de la depuración*

Variable	**VIF**
RCD	∞
CTNCT	∞
DLPC	2.25×10^{15}
T	1.93×10^{7}
LC	1.61×10^{5}
ME	4.58×10^{4}
ROE	2.16×10^{2}
RA	11.61

A partir del dendrograma inicial, se definieron 17 grupos de variables correlacionadas, y dentro de cada uno se seleccionó la variable con menor VIF como representante. Esta estrategia permitió reducir la dimensionalidad manteniendo diversidad temática entre indicadores (liquidez, rentabilidad, eficiencia, cobertura y apalancamiento). En la Tabla 4.3 se presenta el listado final de variables seleccionadas y sus VIF validados.

Tabla 4.3: *Variables seleccionadas por grupo tras evaluación de VIF*

Variable	VIF Final
LG	1.03
T	1.02
RCID	1.02
RCP	1.00
RA	1.00
RCC	1.00
KTNO	1.00
RAF	1.00
PPI	1.00
ROE	1.00
MO	1.00
MB	1.00
AFT	1.00
RI	1.00
PPP	1.00
RAO	1.00
ROI	1.00

Para verificar la efectividad del proceso de depuración, se construyó un nuevo dendrograma con las variables seleccionadas. Como puede observarse en la Figura 4.2, la estructura resultante presenta distancias más elevadas y grupos menos densos, lo cual confirma la reducción de colinealidad en el conjunto de variables finales.

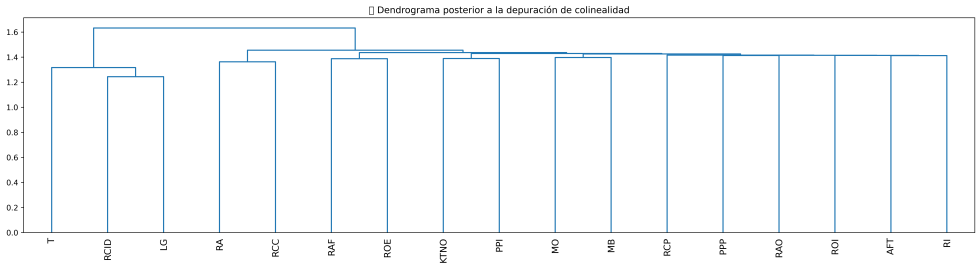

Figura 4.2: Dendrograma posterior a la depuración de colinealidad.

4.4.1. Exploración de estructura latente mediante PCA

Con el objetivo de explorar la estructura latente de los indicadores financieros y evaluar la posibilidad de redundancias estructurales, se aplicó un Análisis de Componentes Principales (PCA) sobre dos versiones de la base de datos: la base completa con 45 indicadores financieros, y la base reducida con las 17 variables seleccionadas tras el control de colinealidad. Aunque el PCA no será utilizado para reducir dimensionalidad en los modelos predictivos, este análisis permite identificar la distribución de la varianza y confirmar empíricamente la existencia de dependencias lineales entre indicadores.

Primero, se aplicó el PCA a la base completa. La Figura 4.3 muestra que los primeros cinco componentes principales explican más del 50 % de la varianza total, y que con diez componentes se alcanza un acumulado superior al 70 %. En particular, el primer componente por sí solo explica más del 18 % de la varianza, lo que indica una fuerte concentración de la información en unos pocos ejes. Esta distribución sugiere una estructura interna altamente redundante, lo cual valida la necesidad de aplicar una reducción por agrupación.

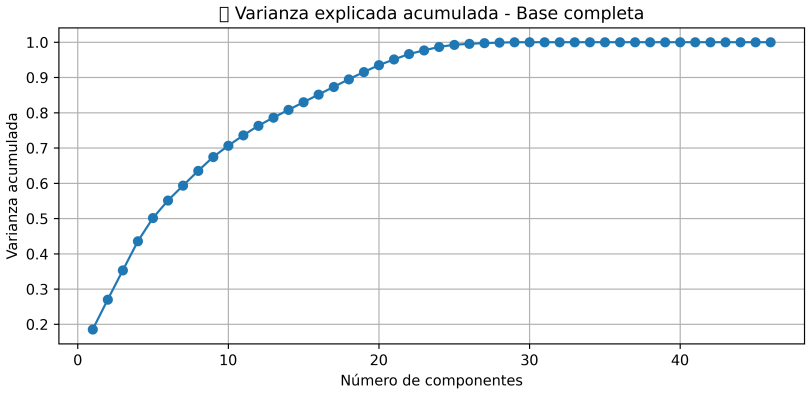

Figura 4.3: Curva de varianza explicada acumulada por PCA en la base completa.

Posteriormente, se aplicó el mismo procedimiento sobre la base reducida de 17 variables seleccionadas. Como se observa en la Figura 4.4, la varianza está más distribuida entre los componentes. Cada uno explica aproximadamente entre 5 % y 7 % de la varianza total, sin que exista un eje dominante. Esta distribución más homogénea refuerza el resultado de la selección, mostrando

que las variables remanentes capturan dimensiones diferentes del fenómeno sin redundancias significativas.

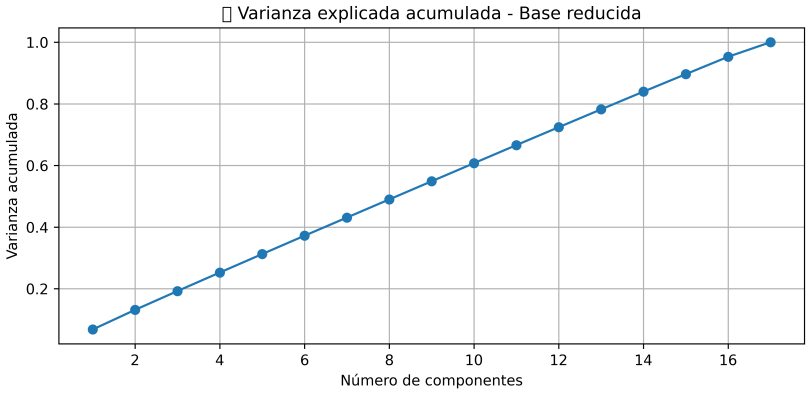

Figura 4.4: Curva de varianza explicada acumulada por PCA en la base reducida.

En conjunto, estos resultados muestran que el conjunto original contenía redundancias importantes, lo que justificó su reducción. Asimismo, evidencian que el conjunto depurado es estructuralmente más equilibrado, adecuado para su uso en modelos predictivos sin incurrir en sobreajuste ni pérdida de interpretabilidad.

Dado que la base reducida de 17 indicadores conserva la diversidad temática original, reduce significativamente la colinealidad y presenta una estructura latente más equilibrada, se realizó una evaluación preliminar de su impacto utilizando el modelo k-NN funcional propuesto. Esta prueba comparativa entre la base completa y la reducida mostró que las diferencias en las métricas de desempeño eran mínimas, lo que valida empíricamente el proceso de depuración. Por lo tanto, y con el objetivo de garantizar eficiencia y comparabilidad metodológica, se optó por utilizar exclusivamente la base reducida en el entrenamiento de los modelos predictivos desarrollados en capítulos posteriores. El análisis detallado de esta evaluación se presenta en el Capítulo 5.

Capítulo 5

Evaluación empírica del modelo funcional

Este capítulo presenta la validación empírica del clasificador funcional k-NN propuesto, aplicado a trayectorias financieras de empresas del sector turismo en Colombia. A partir de una representación funcional construida mediante ventanas móviles de longitud $n = 5$, se implementa un clasificador basado en la distancia entre trayectorias, utilizando una métrica personalizada que penaliza trayectorias incompletas y acota los valores extremos. La variable objetivo es RQ, que indica si la empresa presenta señales de riesgo en su último año reportado.

Con el fin de preservar una comparación metodológicamente consistente con otros modelos predictivos —los cuales operan sobre estructuras tabulares estáticas—, la evaluación principal se realizó exclusivamente sobre trayectorias financieras, sin incluir variables estructurales adicionales. Esta versión del modelo se utiliza como base comparativa frente a modelos como regresión logística, XGBoost o Random Forest.

No obstante, al final del capítulo se presenta una versión extendida del modelo funcional, en la cual se integran variables cualitativas adicionales (como el código sectorial CIIU, el departamento geográfico y el desfase temporal entre trayectorias), con el objetivo de explorar el potencial de mejora y adaptabilidad de la métrica propuesta.

5.1. Modelo funcional para comparación con enfoques tradicionales

Con el objetivo de reducir la dimensionalidad y evitar multicolinealidad entre indicadores, se construyó una base funcional reducida mediante un proceso combinado de agrupamiento jerárquico y selección de variables basado en el Factor de Inflación de la Varianza (VIF). La implementación fue desarrollada íntegramente en Python y ejecutada en Google Colab.

El modelo funcional fue evaluado exclusivamente sobre esta base reducida, utilizando validación cruzada estratificada de 10 folds y una matriz de distancias funcionales previamente calculada. La optimización de hiperparámetros se realizó en dos etapas: primero se ajustaron k (número de vecinos) y λ (penalización por pérdida de datos), y posteriormente se optimizaron los pesos relativos de cada indicador financiero mediante búsqueda bayesiana con `Optuna`.

Las secciones siguientes presentan los resultados obtenidos con los parámetros óptimos, junto con un análisis de importancia relativa de los indicadores financieros. Se incluyen además dos evaluaciones de robustez: (i) una curva de aprendizaje que analiza el desempeño del modelo en distintos tamaños muestrales, y (ii) un procedimiento de bootstrap con 30 repeticiones para estimar la estabilidad esperada de las métricas.

Finalmente, se realiza un análisis aplicado del desempeño del modelo, incluyendo visualizaciones geográficas y sectoriales (tipo mapa de calor) que permiten identificar regiones o sectores con mayor proporción de errores de predicción, aportando elementos útiles para el diagnóstico y la toma de decisiones basada en el modelo.

5.1.1. Optimización de hiperparámetros y pesos por indicador

Con el fin de mejorar tanto el desempeño como la interpretabilidad del clasificador funcional, se realizó un proceso de optimización por etapas utilizando `Optuna`, un framework de optimización bayesiana. La búsqueda se dividió en dos fases: primero se ajustaron los hiperparámetros k (número de vecinos) y λ (penalización por años faltantes), y posteriormente se optimizaron los pesos relativos de cada indicador financiero incluidos en la métrica funcional.

Optimización de k y λ

Se aplicó validación cruzada estratificada de 5 folds sobre una muestra aleatoria del 10 % de la base funcional completa. El objetivo fue maximizar el F1-score promedio. Los rangos explorados fueron:

- $k \in [3, 15]$
- $\lambda \in [0.1, 5.0]$

El mejor desempeño se alcanzó con los siguientes valores:

- $k^* = \mathbf{15}$
- $\lambda^* = \mathbf{1.9915}$
- F1-score promedio = **0.9147**

Optimización de pesos por indicador

A partir de los valores óptimos de k y λ, se procedió a optimizar los pesos asignados a cada dimensión contable. Esta etapa no solo busca una mejora marginal en el desempeño predictivo, sino también proporcionar un criterio empírico para interpretar la importancia relativa de cada indicador dentro de la métrica funcional.

Se utilizó nuevamente validación cruzada de 5 folds sobre una muestra del 10 %. Cada peso w_j fue restringido al intervalo $[0.1, 5.0]$, y posteriormente normalizado para mantener $\sum w_j = 1$.

Los tres indicadores con mayor peso en la distancia funcional optimizada fueron:

- **RAO**: 4.9635
- **RCC**: 4.8677
- **ROI**: 4.7665

El F1-score promedio final, con todos los pesos ajustados, fue de **0.9147**, lo que representa una mejora respecto al modelo con pesos iguales. Sin embargo, más allá de la ganancia marginal en desempeño, el valor principal de este ajuste fino radica en su capacidad de hacer visible la estructura interna de decisión del modelo.

En modelos como el k-NN funcional, no existen coeficientes como en la regresión, por lo que la única forma de identificar qué dimensiones influyen más en las decisiones del modelo es a través de los pesos asignados en la métrica de similitud. Estos pesos, al ser ajustados sobre múltiples folds, constituyen una medida robusta de importancia relativa, diferenciando este enfoque de versiones convencionales del k-NN donde todos los atributos son tratados de forma homogénea.

5.1.2. Evaluación del modelo funcional base con parámetros óptimos

A partir de los valores óptimos encontrados ($k = 15$, $\lambda = 1{,}99$, y pesos diferenciados por indicador), se procedió a ejecutar una evaluación integral sobre el 100 % de la base funcional turística. La validación se realizó mediante validación cruzada estratificada de 10 folds, considerando las métricas clásicas (accuracy, precisión, recall, F1-score, AUC), así como el log loss y la curva Precision–Recall promedio.

Tabla 5.1: *Desempeño del modelo funcional con parámetros optimizados (base completa)*

Métrica	**Promedio**	**Desviación estándar**
Accuracy	0.7682	±0.0136
Precision	0.7723	±0.0119
Recall	0.9446	±0.0125
F1-score	0.8497	±0.0081
AUC	0.8474	±0.0220
Log Loss	0.8210	±0.1664

A continuación se presenta un análisis cualitativo de las métricas reportadas en la Tabla 5.1:

- **Accuracy (76.82 %)**: el modelo acierta en promedio en más del 76 % de los casos. Aunque no es la mejor métrica en contextos desbalanceados, ofrece una visión general del rendimiento total.
- **Precision (77.23 %)**: indica que, de todas las empresas clasificadas como en riesgo, el 77 % realmente lo estaban. Este nivel de precisión es adecuado y permite reducir el número de falsos positivos.
- **Recall (94.46 %)**: destaca por su valor elevado, lo que significa que el modelo logra identificar correctamente a la mayoría de las empresas en riesgo, minimizando los falsos negativos.
- **F1-score (84.97 %)**: al balancear precisión y recall, el F1-score refleja un buen rendimiento global del modelo funcional. La baja desviación sugiere estabilidad.

- **AUC (84.74 %)**: muestra una buena capacidad de discriminación del modelo entre empresas sanas y en riesgo, a través de distintos umbrales de decisión.
- **Log Loss (0.8210)**: indica una penalización moderada por predicciones incorrectas con alta confianza. Esto es esperable en modelos como el k-NN funcional, donde las probabilidades estimadas no están calibradas.

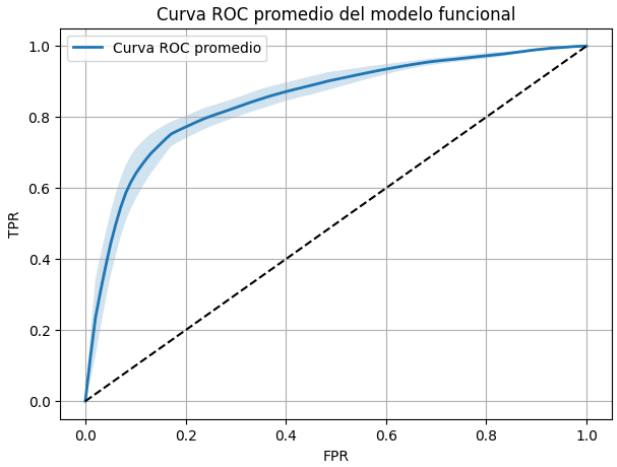

Figura 5.1: Curva ROC promedio del modelo funcional con parámetros optimizados.

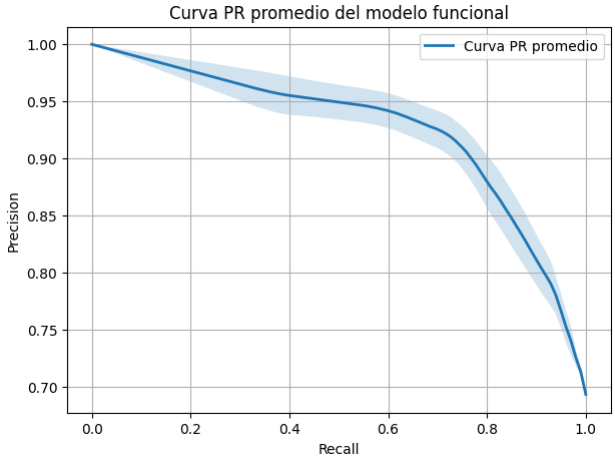

Figura 5.2: Curva Precision–Recall promedio del modelo funcional con parámetros optimizados.

Las curvas ROC y Precision–Recall muestran un buen poder discriminativo, en especial en la recuperación de casos positivos (empresas en riesgo). La banda de ± 1 desviación estándar indica una variabilidad moderada, en línea con lo observado en el F1-score.

Los tres indicadores con mayor peso en la distancia funcional optimizada fueron:

- **RAO** (peso normalizado: 11.1 %)
- **RCC** (peso normalizado: 10.8 %)
- **ROI** (peso normalizado: 10.6 %)

En conjunto, los resultados muestran un **alto desempeño en recall** (detección de riesgo), con una **variabilidad moderada** entre folds. La principal ventaja de este enfoque no solo reside en su precisión, sino también en la posibilidad de interpretar la importancia relativa de cada indicador financiero. Esta capacidad analítica distingue al modelo funcional optimizado frente a versiones tradicionales del k-NN, que no permiten descomponer la distancia en sus componentes informativos.

5.1.3. Robustez del modelo

Curva de aprendizaje del modelo funcional

Con el objetivo de evaluar la robustez del clasificador funcional frente a distintos tamaños muestrales, se construyó una curva de aprendizaje mediante muestras aleatorias estratificadas de proporciones decrecientes del total de empresas. Para cada tamaño muestral se aplicó validación cruzada de 10 folds, manteniendo fijos los hiperparámetros previamente optimizados ($k = 13$, $\lambda = 2.7932$) y los pesos w_j diferenciados por indicador financiero.

La Figura 5.3 presenta los valores promedio del F1-score junto con su desviación estándar para tamaños de muestra desde el 100 % hasta el 1 %. Se observa que el modelo obtiene un F1-score elevado incluso con fracciones muy reducidas de datos, aunque con mayor variabilidad estadística. La curva presenta una ligera disminución en el F1-score a medida que se incrementa el tamaño de la muestra, lo cual puede atribuirse a una pérdida del sobreajuste localizado presente en conjuntos pequeños. Esta tendencia sugiere una transición desde un comportamiento sobreajustado hacia una mayor generalización conforme aumenta la muestra.

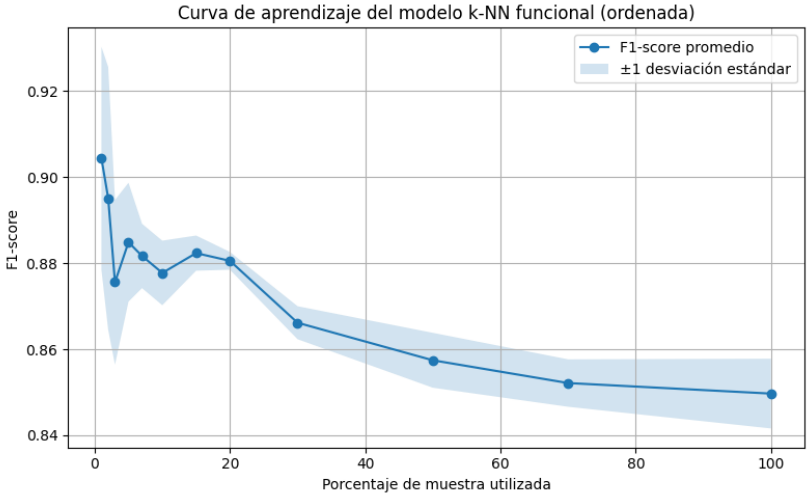

Figura 5.3: Curva de aprendizaje del modelo funcional con métrica personalizada.

Tabla 5.2: *Resultados de la curva de aprendizaje del modelo funcional con métrica personalizada.*

Porcentaje de muestra	F1-score promedio	Desviación estándar
100	0.8511	0.0059
70	0.8529	0.0018
50	0.8651	0.0025
30	0.8619	0.0082
20	0.8783	0.0099
15	0.8789	0.0068
10	0.8784	0.0071
7	0.8887	0.0245
5	0.8787	0.0099
3	0.8984	0.0254
2	0.8631	0.0338
1	0.9002	0.0195

Aunque los F1-scores alcanzan valores cercanos a 0.90 en los extremos inferiores (1 % y 3 %), estos se asocian con desviaciones estándar elevadas, lo que revela una alta sensibilidad a la selección muestral y posibles indicios de sobreajuste. En contraste, a partir del 20 % la curva se estabiliza con F1-scores superiores a 0.86 y desviaciones sistemáticamente bajas. Este rango representa un punto de equilibrio entre rendimiento predictivo, estabilidad estadística y eficiencia computacional.

Estos resultados son particularmente útiles para contextos donde el acceso a la información es limitado, ya que el modelo conserva un rendimiento robusto aun con fracciones reducidas de la base. Para los análisis posteriores, se recomienda utilizar entre el 20 % y el 30 % de la muestra, lo cual asegura una buena capacidad predictiva sin incurrir en costos computacionales innecesarios.

Evaluación de estabilidad mediante *bootstrap*

Para evaluar la estabilidad del modelo funcional frente a la variabilidad de la muestra, se implementó un procedimiento de *bootstrap* mediante submuestreo aleatorio con reemplazo. Se generaron $B = 30$ muestras independientes y, en cada iteración, se aplicó validación cruzada estratificada con 10 folds. Se calcularon métricas clásicas de desempeño manteniendo fijos los hiperparámetros

óptimos ($k = 13$, $\lambda = 2.7932$) y los pesos w_j por indicador financiero.

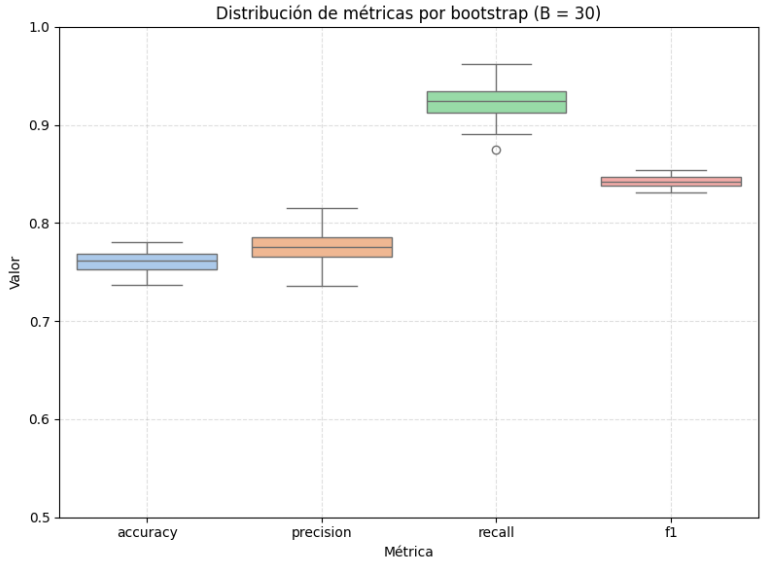

Figura 5.4: Distribución de métricas por *bootstrap* de submuestreo ($B = 30$).

La Figura 5.4 muestra la distribución obtenida para las principales métricas. Se observa una baja dispersión en el F1-score y el recall, lo que indica que el modelo mantiene un rendimiento consistente incluso ante variaciones aleatorias en la muestra. Las métricas promedio y sus desviaciones fueron:

- **Accuracy:** 0.7609 ± 0.0120
- **Precision:** 0.7756 ± 0.0179
- **Recall:** 0.9230 ± 0.0175
- **F1-score:** 0.8426 ± 0.0066
- **AUC:** 0.8284 ± 0.0116
- **Log Loss:** 1.0839 ± 0.1157

Estos resultados confirman la robustez del modelo frente a fluctuaciones en la muestra. En particular, el alto recall con baja variabilidad indica una alta capacidad de detección de empresas en riesgo, mientras que el F1-score refleja

un balance sólido entre precisión y sensibilidad. La estabilidad observada refuerza la confiabilidad del modelo en contextos aplicados.

Además, en todas las iteraciones se mantuvieron como variables de mayor peso en la métrica funcional las siguientes:

- **PPP** (Utilidades sobre patrimonio)
- **ROI** (Retorno sobre la inversión)
- **RCID** (Cobertura de intereses)

Estas variables no solo son relevantes desde el punto de vista financiero, sino que demostraron una alta capacidad discriminativa en el proceso funcional, lo que respalda su utilidad como instrumentos de interpretación y monitoreo.

En síntesis, la evidencia empírica proveniente tanto de la curva de aprendizaje como del procedimiento *bootstrap* respalda la conclusión de que el modelo funcional propuesto presenta una *alta robustez*, una capacidad predictiva consistente y una estructura estable de variables explicativas. Estas propiedades lo convierten en una herramienta sólida para la evaluación de riesgo empresarial, especialmente útil en entornos con restricciones de datos o características heterogéneas.

5.1.4. Análisis aplicado del modelo funcional

Esta sección presenta tres análisis complementarios que permiten interpretar con mayor profundidad el comportamiento del clasificador k-NN funcional optimizado. Se incluyen visualizaciones por sector económico, por ubicación territorial y por la importancia relativa de los indicadores financieros utilizados en la métrica funcional.

Desempeño por sector económico

La Figura 5.5 muestra la tasa promedio de error por sector económico, agrupados según la letra CIIU y posteriormente renombrados con una descripción simplificada. Los sectores con mayor tasa de error relativa fueron *Servicios administrativos* (51 %), *Alojamiento y comida* (33 %) y *Comercio y vehículos* (32 %). En contraste, los sectores con mejor desempeño correspondieron a *Información y comunicaciones* (12 %) y *Finanzas y seguros* (20 %).

Figura 5.5: Tasa de error promedio por sector económico (*elaboración propia*)

Desempeño territorial por departamento

La Figura 5.6 presenta la tasa de error por departamento, considerando únicamente aquellos con al menos 20 empresas en la base funcional final. Los errores más bajos se observaron en departamentos como Boyacá (18 %), Meta (18 %) y Caldas (19 %), mientras que Cesar (32 %), Magdalena (30 %) y San Andrés y Providencia (30 %) registraron los valores más altos.

Figura 5.6: Tasa de error promedio por departamento (*elaboración propia*)

5.1.5. Importancia relativa de los indicadores financieros

La Figura 5.7 presenta un gráfico tipo radar con los pesos normalizados asignados a cada indicador financiero durante el proceso de optimización del modelo. Este gráfico permite visualizar qué dimensiones contables tuvieron mayor influencia en la distancia funcional final.

Se destacan como variables más relevantes: `RAO` (Rendimiento del Activo Operacional), `RCC` (Razón de Cobertura de Cargos) y `ROI` (Retorno sobre la Inversión), todas con pesos superiores al 10 %.

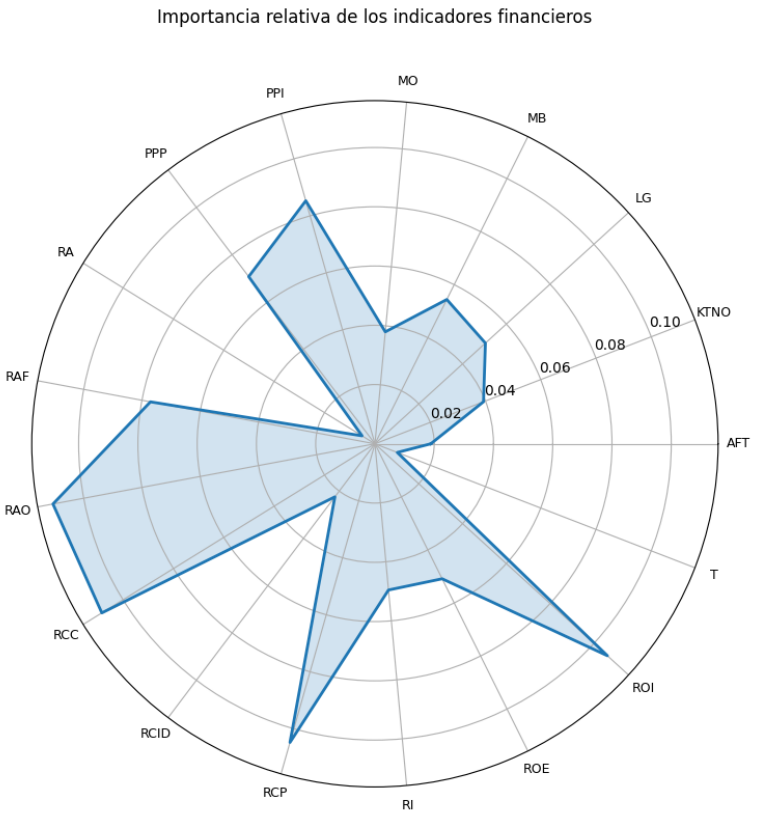

Figura 5.7: Importancia relativa de los indicadores financieros (*elaboración propia*)

El ajuste fino de los pesos por indicador no tiene como único objetivo mejorar el desempeño del modelo, sino, sobre todo, aportar interpretabilidad funcional. A diferencia de modelos como la regresión, el k-NN funcional no ofrece coeficientes directamente interpretables. Por ello, los pesos asignados en la métrica de similitud constituyen la principal vía para entender qué dimensiones financieras influyen más en las decisiones del clasificador. Esta capacidad de interpretación representa una ventaja importante frente a modelos más opacos, y fortalece la utilidad analítica del enfoque funcional.

5.2. Versión final del modelo funcional con variables categóricas y temporales

Esta sección presenta la versión final del modelo funcional propuesto, en la cual se implementa una métrica de distancia compuesta que amplía el enfoque original basado exclusivamente en trayectorias financieras. La nueva métrica incorpora tres dimensiones adicionales: el departamento geográfico (`DEP`), el sector económico principal según el código CIIU, y el desfase temporal entre trayectorias (diferencia entre los años finales de las ventanas móviles).

Las dos primeras dimensiones corresponden a variables categóricas estáticas que permiten capturar similitudes estructurales o institucionales entre empresas. La tercera es una variable cuantitativa discreta que introduce una noción de sincronización relativa entre trayectorias financieras. Este componente busca compensar, al menos parcialmente, el desalineamiento cronológico entre empresas, lo cual resulta relevante para identificar patrones recientes frente a contextos más rezagados. Aunque este desfase no modela directamente eventos exógenos (como la pandemia del COVID-19 o crisis sectoriales), sí refleja el grado de contemporaneidad entre los datos comparados.

La métrica extendida mantiene como núcleo el componente funcional previamente definido, pero ahora se complementa con términos que capturan explícitamente la heterogeneidad sectorial, territorial y temporal. Cada uno de estos componentes fue integrado de forma explícita en la métrica global, asignándole un peso específico ajustado mediante un proceso de optimización conjunta. La implementación replicó los pasos descritos en capítulos anteriores: construcción del espacio funcional, optimización de los hiperparámetros k y λ, y ajuste de los pesos por indicador financiero, por variable categórica y por desfase temporal.

La versión final presentada aquí corresponde a la incorporación formal de las extensiones **1** (ver Subsección 3.4.1) y **3** (ver Subsección 3.4.3) discutidas en el Capítulo 3, y constituye la forma más completa y robusta del modelo desarrollado en esta tesis.

No obstante, para garantizar una comparación metodológicamente consistente con modelos clásicos como regresión logística, XGBoost o Random Forest —los cuales operan sobre estructuras tabulares estáticas y no permiten la incorporación directa del desfase temporal entre trayectorias—, el análisis comparativo del siguiente capítulo se realizó utilizando únicamente la versión base del modelo funcional.

5.2.1. Redefinición de la métrica con variables categóricas y temporales

Esta sección presenta la extensión formal de la métrica de distancia utilizada, incorporando variables categóricas y cuantitativas estáticas dentro del cálculo de distancias entre empresas. El espacio de representación final se define como el producto cartesiano:

$$\mathcal{X} = \mathcal{S} \times \mathcal{Q} \times \mathcal{F}$$

donde:

- $\mathcal{S}$ representa el espacio de variables categóricas estáticas, en este caso, el Departamento geográfico y el Sector económico,
- $\mathcal{Q}$ representa el espacio de variables cuantitativas estáticas, como el desfase temporal entre trayectorias,
- $\mathcal{F}$ representa el espacio funcional multivariado definido por trayectorias financieras a lo largo del tiempo.

A continuación, se detallan los tres componentes de la métrica:

1. Componente categórica estática ($\mathcal{S}$) Las variables categóricas se comparan utilizando la métrica discreta. En este trabajo se adopta una forma simple y directa de esta métrica, útil para capturar similitudes estructurales sin necesidad de codificación adicional.

Sea y_1^e y y_2^e el vector de variables categóricas para la empresa e, la distancia categórica se define como:

$$D_{\mathcal{S}}(x^e, x^{e'}) = \delta_1 + \delta_2$$

donde:

$$\delta_j = \begin{cases} 0 & \text{si } y_j^e = y_j^{e'} \\ \alpha_j & \text{si } y_j^e \neq y_j^{e'} \end{cases} \quad \text{para } j = 1, 2$$

Aquí, α_j representa el peso asignado a la discrepancia en la categoría j, y se ajusta mediante el proceso de optimización conjunta del modelo.

2. Componente cuantitativa estática ($\mathcal{Q}$) Este componente mide la diferencia absoluta entre los años finales observados en las trayectorias funcionales de cada empresa, es decir, el desfase temporal entre las ventanas móviles utilizadas para representar sus indicadores financieros. Esta diferencia busca capturar el grado de sincronización relativa entre empresas: trayectorias más contemporáneas pueden compartir patrones recientes, mientras que trayectorias más antiguas pueden estar desfasadas frente a shocks recientes o cambios estructurales.

Para calcular esta discrepancia se utiliza la métrica estándar sobre los números reales, que en este caso corresponde a la distancia euclidiana en $\mathbb{R}$.

Formalmente, se define como:

$$D_{\mathcal{Q}}(x^e, x^{e'}) = \beta_3 \cdot \left| y_3^e - y_3^{e'} \right|$$

donde:

- y_3^e representa el año final de observación para la empresa e,
- β_3 es el peso asignado a esta discrepancia temporal,
- la diferencia absoluta $\left| y_3^e - y_3^{e'} \right|$ refleja el grado de desalineamiento cronológico entre ambas trayectorias.

3. Componente funcional ($\mathcal{F}$) Se mantiene la distancia funcional previamente definida, basada en trayectorias temporales multivariadas con penalización por pérdida de dominio y transformación acotada. Para cada indicador financiero $j = 1, \ldots, r$, la distancia se calcula como:

$$D_{\mathcal{F}}(x^e, x^{e'}) = \sum_{j=1}^{r} w_j \cdot \left(\frac{\sum_{t=k_j}^{0} \left| x_{j,t}^e - x_{j,t}^{e'} \right| \cdot \left(1 + \lambda \cdot \frac{k_j + t - 1}{\ell} \right)}{1 + \sum_{t=k_j}^{0} \left| x_{j,t}^e - x_{j,t}^{e'} \right| \cdot \left(1 + \lambda \cdot \frac{k_j + t - 1}{\ell} \right)} \right)$$

donde:

- $x_{j,t}^e$: valor del indicador j para la empresa e en el año t,
- w_j: peso asignado al indicador j,
- λ: parámetro de penalización temporal,

- k_j: año más antiguo con datos para el indicador j,
- ℓ: longitud de la ventana temporal (por ejemplo, $\ell = 5$).

Distancia total combinada. La métrica compuesta final entre dos empresas se expresa como:

$$D(x^e, x^{e'}) = \underbrace{\sum_{j=1}^{n_c} \alpha_j \cdot d_j^{(c)}(y_j^{(c),e}, y_j^{(c),e'})}_{\text{Categóricas}} + \underbrace{\sum_{j=1}^{n_q} \beta_j \cdot d_j^{(q)}(y_j^{(q),e}, y_j^{(q),e'})}_{\text{Cuantitativas}} + \underbrace{\sum_{j=1}^{r} w_j \cdot \left(\frac{\sum_{t=k_j}^{0} \left| x_{j,t}^{e} - x_{j,t}^{e'} \right| \cdot \left(1 + \lambda \cdot \frac{k_j + t - 1}{\ell}\right)}{1 + \sum_{t=k_j}^{0} \left| x_{j,t}^{e} - x_{j,t}^{e'} \right| \cdot \left(1 + \lambda \cdot \frac{k_j + t - 1}{\ell}\right)} \right)}_{\text{Funcional}} \quad (5.2.1)$$

En esta expresión, $d_j^{(c)}(\cdot, \cdot)$ representa una métrica definida sobre el espacio categórico para la variable j, que en este caso corresponde a la métrica discreta para ambas categorías utilizadas. Por su parte, $d_j^{(q)}(\cdot, \cdot)$ es una métrica aplicada a variables cuantitativas, correspondiente aquí a la métrica estándar sobre $\mathbb{R}$ (distancia euclidiana en dimensión uno). Finalmente, el tercer componente mantiene la estructura funcional previamente desarrollada, que incorpora penalización por pérdida de dominio, transformación acotada y combinación ponderada entre indicadores. La suma ponderada de estos tres términos permite integrar de manera coherente distintas fuentes de heterogeneidad estructural, temporal y financiera en el cálculo de la distancia total.

5.2.2. Optimización de parámetros del modelo final

El ajuste de la métrica extendida se llevó a cabo en dos etapas, utilizando validación cruzada estratificada y búsqueda bayesiana con `Optuna`. La optimización se realizó sobre muestras aleatorias del 10 % del conjunto funcional reducido, maximizando el F1-score promedio del clasificador.

En una primera etapa se ajustaron los hiperparámetros k (número de vecinos) y λ (penalización por pérdida de dominio funcional), obteniendo como valores óptimos $k = 3$ y $\lambda = 4{,}3443$. En la segunda etapa, con estos valores fijos, se optimizaron conjuntamente los pesos asignados a cada componente de la métrica: los indicadores financieros (w_j), las variables categóricas (α_1 y α_2) y el desfase temporal (β_3).

El modelo alcanzó un F1-score promedio de 0,9147 tras la optimización completa. Las tres dimensiones con mayor peso relativo dentro de la métrica fueron el desfase temporal ($\beta_3 = 4{,}8427$), el indicador financiero AFT ($w_{\text{AFT}} = 4{,}5561$) y el indicador PPI ($w_{\text{PPI}} = 4{,}3598$). Estos resultados confirman que la incorporación de componentes estructurales y temporales no solo mejora la precisión del modelo, sino que también aporta interpretabilidad al distinguir las fuentes de similitud más influyentes entre empresas.

5.2.3. Evaluación del modelo funcional extendido

Una vez definidos los parámetros óptimos del modelo funcional extendido ($k = 3$, $\lambda = 4{,}3443$, y pesos diferenciados por variable), se ejecutó una evaluación integral sobre el 100 % de la base funcional. Se utilizó validación cruzada estratificada de 10 folds, considerando métricas clásicas de desempeño y medidas adicionales de robustez.

Tabla 5.3: *Desempeño del modelo funcional extendido con parámetros óptimos*

Métrica	**Promedio**	**Desviación estándar**
Accuracy	0.9730	±0.0064
Precision	0.9976	±0.0019
Recall	0.9635	±0.0086
F1-score	0.9802	±0.0048
AUC	0.9854	±0.0051
Log Loss	0.6860	±0.2272

El modelo mostró un desempeño sobresaliente, destacándose especialmente en F1-score (98.02 %) y AUC (98.54 %), con baja variabilidad entre folds. La curva de aprendizaje (Figura 5.8) mostró que el rendimiento se estabiliza rápidamente con muestras pequeñas, mientras que la distribución por bootstrap (Figura 5.9) confirmó la robustez del modelo, con métricas consistentemente altas en distintas particiones.

Los tres componentes con mayor peso relativo dentro de la métrica fueron el desfase temporal (peso normalizado: 9.2 %), seguido por los indicadores financieros AFT (8.6 %) y PPI (8.2 %). Estos resultados refuerzan la relevancia del componente temporal como fuente principal de similitud predictiva en contextos financieros heterogéneos.

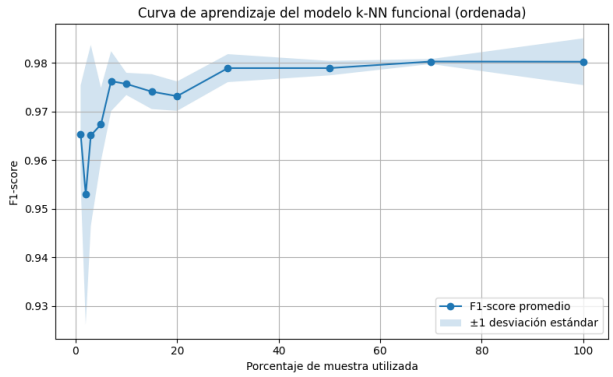

Figura 5.8: Curva de aprendizaje del modelo funcional con parámetros extendidos.

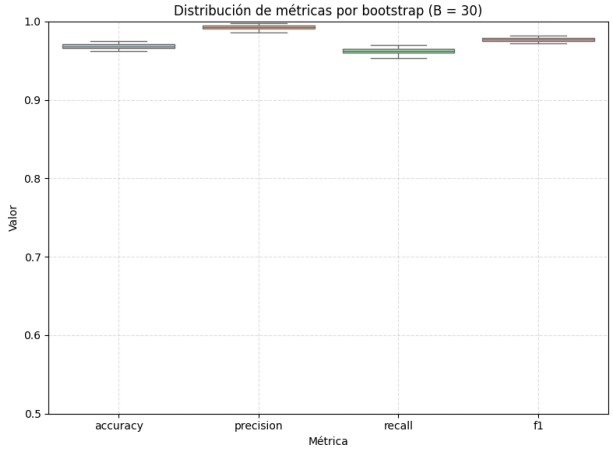

Figura 5.9: Distribución de métricas por bootstrap (modelo funcional extendido).

5.3. Reproducibilidad del modelo funcional y disponibilidad del código

Toda la implementación computacional del modelo funcional propuesto en esta tesis ha sido desarrollada en lenguaje Python utilizando Google Colab como entorno principal. Los scripts abarcan el procesamiento de datos, construcción del espacio funcional, definición de la métrica personalizada (en sus versiones base y extendida), validación cruzada, optimización de hiperparámetros mediante `Optuna`, y generación automatizada de gráficos y reportes.

Con el fin de garantizar la trazabilidad y reproducibilidad del modelo, se ha dispuesto un repositorio público en GitHub que contiene todos los archivos necesarios para replicar los resultados presentados. Dicho repositorio incluye una base mínima de entrada y los scripts organizados por etapas, siguiendo la lógica del flujo metodológico descrito en esta tesis.

- **Repositorio GitHub:**
  https://github.com/tesisluisruizparedes/Tesis_KNN_Funcional
- **Carpeta `Datos/`:** base mínima funcional con trayectorias financieras y variables estructurales, contenida en el archivo
  `8_Base_Funcional_Reducida_RQ.parquet`
- **Carpeta `scripts/`:** scripts completos del modelo funcional, incluyendo la versión extendida con variables categóricas y temporales
- **Carpeta `scripts_comparativos/`:** implementación de todos los modelos alternativos evaluados en esta tesis,
  incluyendo regresiones, árboles, métodos secuenciales e híbridos

Adicionalmente, el repositorio cuenta con enlaces directos para ejecutar cada etapa del modelo funcional en Google Colab, facilitando su adaptación por parte de otros investigadores o analistas. Esta organización promueve la transparencia científica y la utilidad práctica del enfoque propuesto.

5.4. Aplicación interactiva del modelo funcional en entorno web

Como parte del desarrollo práctico de esta tesis, se construyó una aplicación web interactiva utilizando `Streamlit`, disponible públicamente en línea[1]. Esta herramienta implementa la versión extendida del modelo funcional k-NN, incorporando tanto trayectorias financieras como variables estructurales (sector económico, región y desfase temporal) dentro del cálculo de similitud.

El usuario puede ingresar los valores observados de los 17 indicadores financieros de una empresa para los últimos cinco años —incluso con datos faltantes—, así como seleccionar su Departamento y sector económico. La aplicación devuelve como salida:

- La probabilidad estimada de riesgo de quiebra, calculada a partir de la proporción de vecinos históricos en riesgo.
- Un listado de las trayectorias más similares (vecinos funcionales), incluyendo el NIT, año final, sector económico y distancia funcional.

Figura 5.10: Interfaz de la aplicación web funcional desarrollada en `Streamlit`.

Esta herramienta fue construida con el propósito de facilitar la evaluación predictiva de nuevas empresas y potenciar la interpretabilidad del modelo. La posibilidad de identificar y visualizar las trayectorias más similares aporta valor práctico al análisis financiero comparado, fortaleciendo así la trazabilidad, utilidad y aplicabilidad del enfoque funcional propuesto.

[1] https://tesis-knn.streamlit.app/

Para preservar la confidencialidad de las trayectorias financieras históricas, la aplicación solo muestra el NIT y algunas variables categóricas (como el departamento y el sector económico) de las empresas más similares. Esta decisión permite mantener la trazabilidad pública, ya que cualquier usuario puede consultar la información completa de dichas empresas a través del sistema SIIS de la Superintendencia de Sociedades[2]. No se incluye razón social ni valores financieros detallados, en concordancia con buenas prácticas de uso ético de datos empresariales.

[2]https://siis.supersociedades.gov.co/

Capítulo 6

Comparación de modelos predictivos

Zhao et al. (2024b) presentan una revisión crítica de 149 estudios sobre predicción de quiebra y distress financiero, destacando la evolución del campo desde los modelos estadísticos tradicionales hasta técnicas avanzadas de inteligencia artificial. Complementariamente, Dasilas y Rigani (2024) realizan una revisión sistemática de 207 estudios centrados en el uso de técnicas de aprendizaje automático para predicción de quiebra entre 2012 y 2023, resaltando el creciente protagonismo de modelos híbridos y de estrategias específicas para tratar datos desbalanceados. Esta progresión ha dado lugar a una diversidad de enfoques que varían en complejidad, tipo de datos utilizados y criterios de evaluación, lo cual refuerza la necesidad de comparar sistemáticamente distintos modelos bajo condiciones similares.

Como parte del análisis comparativo, se implementaron modelos predictivos con distintos niveles de complejidad, naturaleza y enfoque temporal. El objetivo principal fue evaluar el desempeño relativo de cada uno frente a la tarea de predecir riesgo de quiebra en empresas del sector turismo, usando como insumo la misma base de datos. Aljawazneh et al. (2021) destacan precisamente la importancia de realizar comparaciones integrales entre modelos clásicos, técnicas de ensamblaje y métodos de deep learning, especialmente en contextos con datos financieros desbalanceados, donde las combinaciones con técnicas de remuestreo como SMOTE-ENN pueden marcar diferencias sustantivas en el rendimiento. De manera complementaria, Antulov-Fantulin et al. (2021) subrayan que, más allá del algoritmo utilizado, la inclusión de variables no financieras —como la localización geográfica o factores institucionales—

puede mejorar significativamente la capacidad predictiva de los modelos de quiebra, lo que refuerza la necesidad de evaluaciones amplias y contextuales. En esa misma línea, Zhao et al. (2024a) proponen la incorporación de relaciones estructurales entre empresas mediante análisis de redes complejas y técnicas de representación como Node2Vec, evidenciando que estos nuevos predictores basados en grafos pueden mejorar la precisión en la detección del riesgo de quiebra. Barboza et al. (2021) desarrollan un amplio estudio empírico en el que combinan métodos estadísticos y de aprendizaje automático para construir cientos de modelos, concluyendo que las técnicas modernas superan a los enfoques clásicos, especialmente cuando se incorporan variables dinámicas como tasas y variaciones de crecimiento. Finalmente, Barboza et al. (2017) comparan ocho técnicas de clasificación aplicadas a empresas de EE.UU. y Canadá, mostrando que los modelos de machine learning —en particular random forest— logran mejoras del 10 % en precisión frente a enfoques tradicionales como la regresión logística o el análisis discriminante.

Se incluyeron modelos tradicionales estáticos como la regresión logística, el modelo Probit y el k-NN clásico, los cuales destacan por su simplicidad e interpretabilidad. También se evaluaron modelos basados en árboles de decisión como Random Forest, XGBoost, LightGBM y CatBoost, tanto en su forma básica como con variantes que incorporan SMOTE para abordar el desbalance de clases.

Adicionalmente, se consideraron modelos funcionales de tipo k-NN que operan sobre trayectorias multivariadas de indicadores financieros, pero utilizando métricas estándar como la distancia L1, L2, Mahalanobis o transformaciones derivadas.

También se incluyeron modelos secuenciales basados en redes neuronales como LSTM, BiLSTM y CNN-LSTM, que capturan la evolución financiera de las empresas a lo largo del tiempo.

Finalmente, se evaluaron modelos híbridos que combinan técnicas supervisadas, redes generativas o ensamblajes de modelos, como XGBoost con redes neuronales, GAN combinadas con LightGBM, y esquemas de *stacking* entre modelos heterogéneos. Estas configuraciones representan algunos de los enfoques más recientes en aprendizaje automático para problemas de clasificación complejos.

Para asegurar una comparación justa entre enfoques heterogéneos, todos los modelos fueron entrenados y evaluados sobre la misma base de entrada, construida a partir de los 17 indicadores financieros seleccionados mediante análisis de agrupamiento y VIF. Esta base reducida no incluye variables

categóricas ni transformaciones derivadas, lo cual permite aislar el efecto del tipo de modelo (estático, funcional, secuencial, híbrido) sobre el desempeño predictivo. El modelo funcional extendido, que incorpora adicionalmente variables categóricas y temporales en la métrica de similitud, fue evaluado por separado sobre una base con estructura más rica. Por tanto, su comparación con los demás modelos debe interpretarse como una evidencia complementaria del valor agregado que supone integrar trayectorias temporales y atributos estructurales en este tipo de enfoques.

Los resultados de cada grupo de modelos se presentan en las secciones siguientes, con un enfoque comparativo que destaca no solo su desempeño numérico, sino también sus ventajas relativas en términos de robustez, interpretabilidad y aplicabilidad práctica.

6.1. Metodología comparativa

Todos los modelos evaluados en este capítulo fueron entrenados y validados utilizando la misma base de datos, compuesta exclusivamente por empresas del sector turismo en Colombia. Esta base fue construida a partir de un conjunto común de 17 indicadores financieros seleccionados mediante análisis exploratorio, agrupamiento jerárquico y verificación de colinealidad.

Para asegurar la comparabilidad entre modelos, se utilizó una única variable objetivo —el indicador binario de riesgo de quiebra (`RQ`)— y se aplicó un protocolo de validación cruzada estratificada de 10 *folds* en todos los casos. De esta forma, cada modelo fue evaluado con los mismos subconjuntos de entrenamiento y prueba, garantizando condiciones equitativas para medir su desempeño.

Se calcularon métricas clásicas de clasificación, incluyendo *accuracy*, *precision*, *recall*, *F1-score*, *AUC* (área bajo la curva ROC) y *log loss*. Los resultados se presentan como promedio y desviación estándar sobre los 10 *folds*, lo que permite evaluar tanto el nivel de desempeño como su estabilidad. En los modelos más relevantes se complementó el análisis con curvas ROC y PRC promedio, así como con visualización de variables importantes según coeficientes, pesos o valores SHAP.

En el caso de modelos afectados por el desbalance de clases, se exploraron variantes con *SMOTE* (Synthetic Minority Over-sampling Technique) para mejorar la capacidad de detección de los casos positivos. Asimismo, en algunos modelos avanzados se aplicaron estrategias de optimización de hiperparámetros utilizando búsqueda bayesiana con `Optuna`, buscando mejorar su ajuste y desempeño general.

Todos los modelos fueron ejecutados sobre la misma estructura de datos: ya sea como vectores estáticos, ventanas dinámicas de cinco años o trayectorias funcionales completas. Esto permite una comparación transversal no solo en términos de precisión, sino también del tipo de información que cada modelo es capaz de incorporar y aprovechar.

6.2. Resultados comparativos por tipo de modelo

A continuación se presentan los resultados obtenidos por cada grupo de modelos evaluados. La comparación se organiza según el tipo de modelo y su complejidad metodológica, con el fin de facilitar el análisis entre enfoques tradicionales, funcionales, avanzados, secuenciales e híbridos. Cada bloque incluye un resumen de métricas relevantes y una breve discusión sobre su desempeño relativo.

6.2.1. Modelos tradicionales estáticos

Como punto de partida, se evaluaron cuatro modelos tradicionales basados en variables estáticas: regresión logística, modelo Probit, k-NN clásico y una variante dinámica denominada *Rolling Logit*. Estos enfoques son ampliamente conocidos por su simplicidad y facilidad de interpretación, y funcionan como línea base para comparar el desempeño de modelos más sofisticados. Aunque existen alternativas probabilísticas más complejas, como los Gaussian Processes propuestos por Antunes et al. (2017), los modelos tradicionales conservan su relevancia como punto de partida debido a su bajo costo computacional y su utilidad como referencia comparativa. Barboza y Altman (2024) confirman esta vigencia al utilizar la regresión logística como benchmark principal en su análisis de distress financiero en empresas latinoamericanas, donde si bien modelos como random forest mostraron mejor desempeño, la regresión permaneció como un punto de referencia indispensable.

Tabla 6.1: *Desempeño promedio de modelos tradicionales estáticos (validación cruzada 10 folds)*

Modelo	Accuracy	Precision	Recall	F1-score	AUC	LogLoss
Regresión logística	0,5894 ± 0,0091	0,1740 ± 0,0071	0,5536 ± 0,0217	0,2648 ± 0,0105	0,5973 ± 0,0153	0,6783 ± 0,0041
Probit	0,8667 ± 0,0004	0,5990 ± 0,1748	0,0066 ± 0,0022	0,0130 ± 0,0043	0,5963 ± 0,0155	0,3885 ± 0,0048
k-NN clásico	0,8599 ± 0,0027	0,4420 ± 0,0223	0,1834 ± 0,0082	0,2591 ± 0,0103	0,6652 ± 0,0063	1,9451 ± 0,0623
Rolling Logit	0,5378 ± 0,0355	0,1595 ± 0,0085	0,6207 ± 0,0578	0,2533 ± 0,0128	0,6072 ± 0,0157	0,7170 ± 0,0284

Fuente: Elaboración propia con base en resultados de validación cruzada estratificada (10 folds).

En esta categoría, el modelo de regresión logística mostró un desempeño aceptable como línea base, con un F1-score de 0,2648 y un *recall* del 55,36 %, lo que indica una razonable capacidad para identificar empresas con riesgo de quiebra. El modelo Probit, en cambio, obtuvo un F1-score de apenas 0,013 a pesar de tener una alta precisión, evidenciando una sensibilidad casi nula al evento de interés.

El modelo k-NN clásico logró un *accuracy* alto (0,8599), pero con un *recall* limitado, lo cual sugiere que favorece la clasificación de empresas sanas en detrimento de la detección de riesgo. Finalmente, el Rolling Logit —aunque incorpora una estructura dinámica simple— presentó un *recall* competitivo (62,07 %), pero un F1-score bajo, reflejando una pobre precisión en sus predicciones positivas.

Estos modelos sirven como referencia fundamental frente a la cual se comparan los enfoques más complejos desarrollados en secciones posteriores.

6.2.2. Modelos k-NN funcionales con métricas estándar

En esta sección se presentan los resultados de distintas versiones del modelo k-NN funcional construidas con métricas de distancia convencionales o adaptadas, distintas a la métrica personalizada propuesta en esta tesis. Todas las variantes emplean trayectorias multivariadas de cinco años y fueron evaluadas con los mismos hiperparámetros ($k = 14$, $\lambda = 9$) para facilitar la comparación entre métodos de distancia.

Tabla 6.2: *Desempeño promedio de modelos k-NN funcionales con métricas estándar (validación cruzada 10 folds)*

Modelo	Accuracy	Precision	Recall	F1-score
k-NN funcional (L1)	0,8068	0,1475	0,3756	0,2098 ± 0,0384
k-NN funcional (L2)	0,7948	0,1296	0,3496	0,1866 ± 0,0206
k-NN funcional (Promedio)	0,8068	0,1475	0,3756	0,2098 ± 0,0384
k-NN funcional (PCA funcional)	0,7947	0,1308	0,3640	0,1921 ± 0,0349
k-NN funcional (Derivadas)	0,8133	0,1503	0,3543	0,2098 ± 0,0569
k-NN funcional (Mahalanobis)	0,7854	0,1358	0,4036	0,2026 ± 0,0498

Fuente: Elaboración propia con base en resultados de validación cruzada estratificada (10 folds). Las métricas AUC y LogLoss no fueron calculadas, ya que estas variantes no generan estimaciones probabilísticas.

Los resultados muestran que, aunque todas las variantes funcionales logran niveles aceptables de *accuracy* (superiores al 78 %), sus F1-score se mantienen en niveles bajos, oscilando entre 0,1866 y 0,2098, con desviaciones relativamente altas. Esto refleja una capacidad limitada para balancear precisión y sensibilidad.

La versión basada en distancia Mahalanobis obtuvo el mayor *recall* (40,36 %), mientras que la versión con derivadas alcanzó el mayor *accuracy* general (81,33 %). Sin embargo, en ningún caso se superó el desempeño obtenido por la métrica funcional personalizada presentada en el capítulo anterior.

Estas variantes permiten evidenciar que la simple representación funcional no garantiza mejoras sustantivas en la predicción si no va acompañada de una métrica adecuada para capturar similitudes reales entre trayectorias empresariales. Esto justifica la necesidad del diseño e integración de una métrica específica, como la desarrollada en esta investigación.

6.2.3. Modelos avanzados basados en árboles de decisión

Esta sección presenta modelos construidos sobre algoritmos de tipo *ensemble* basados en árboles de decisión. Se incluyen tanto versiones estándar como variantes optimizadas con técnicas como SMOTE y búsqueda de hiperparámetros. También se incorporan configuraciones extendidas mediante ventanas temporales apiladas de cinco años por empresa, lo cual permite capturar relaciones temporales manteniendo una estructura de datos tabular. Zieba et al. (2016) destacan que el desempeño de modelos como XGBoost puede potenciarse aún más mediante la generación evolutiva de variables sintéticas, capaces de reflejar relaciones no lineales entre indicadores financieros, lo cual mejora la precisión incluso en contextos con alto desbalance de clases.

Tabla 6.3: *Desempeño promedio de modelos basados en árboles de decisión (validación cruzada 10 folds)*

Modelo	Accuracy	Precision	Recall	F1-score	AUC	LogLoss
Random Forest	0,8714 ± 0,0011	0,6698 ± 0,0455	0,0744 ± 0,0039	0,1339 ± 0,0063	0,7790 ± 0,0088	0,3417 ± 0,0100
Random Forest (SMOTE)	0,8332 ± 0,0059	0,3535 ± 0,0214	0,2989 ± 0,0131	0,3238 ± 0,0158	0,7331 ± 0,0109	0,4255 ± 0,0090
XGBoost	0,8783 ± 0,0026	0,6192 ± 0,0280	0,2309 ± 0,0112	0,3362 ± 0,0143	0,7802 ± 0,0131	0,3264 ± 0,0084
XGBoost (SMOTE)	0,8404 ± 0,0066	0,3989 ± 0,0228	0,3811 ± 0,0137	0,3896 ± 0,0166	0,7496 ± 0,0116	0,3910 ± 0,0091
LightGBM (SMOTE)	0,8216 ± 0,0058	0,3605 ± 0,0146	0,4327 ± 0,0106	0,3932 ± 0,0123	0,7523 ± 0,0114	0,4287 ± 0,0062
Random Forest Dinámico	0,8748 ± 0,0015	0,6104 ± 0,1455	0,0288 ± 0,0083	0,0548 ± 0,0154	0,7308 ± 0,0147	0,3560 ± 0,0155
XGBoost (SMOTE + Dummies + Full Vars)	0,8489 ± 0,0057	0,3744 ± 0,0338	0,1920 ± 0,0104	0,2536 ± 0,0163	0,6985 ± 0,0121	0,3883 ± 0,0080
CatBoost (SMOTE)	0,8689 ± 0,0033	0,5365 ± 0,0473	0,1375 ± 0,0121	0,2188 ± 0,0187	0,7350 ± 0,0117	0,3521 ± 0,0054
XGBoost (Optuna + SMOTE + Año)	0,8863 ± 0,0044	0,6406 ± 0,0308	0,3393 ± 0,0200	0,4434 ± 0,0229	0,8087 ± 0,0098	0,3531 ± 0,0125
LightGBM (Optuna + SMOTE)	0,8871 ± 0,0042	0,6616 ± 0,0347	0,3172 ± 0,0164	0,4287 ± 0,0210	0,8142 ± 0,0111	0,3033 ± 0,0071
XGBoost (ventana 5 años)	0,8397 ± 0,0179	0,9091 ± 0,0147	0,8824 ± 0,0186	0,8954 ± 0,0120	0,8996 ± 0,0109	0,3519 ± 0,0244
LightGBM (ventana 5 años)	0,8476 ± 0,0146	0,9173 ± 0,0118	0,8840 ± 0,0187	0,9002 ± 0,0101	0,9010 ± 0,0106	0,3476 ± 0,0242

Fuente: Elaboración propia con base en resultados de validación cruzada estratificada (10 folds).

Los modelos de árboles de decisión muestran resultados notoriamente superiores a los modelos tradicionales. En particular, las versiones optimizadas como XGBoost con Optuna y SMOTE, y LightGBM con ventana expandida, lograron F1-scores por encima de 0,89 y áreas bajo la curva ROC mayores a 0,90, lo que representa un desempeño predictivo sobresaliente.

La inclusión de ventanas de cinco años permitió capturar dinámicas temporales sin requerir modelos secuenciales, mientras que la combinación con técnicas de optimización o variables adicionales (como el año) reforzó la sensibilidad y precisión.

Este conjunto de modelos destaca por su capacidad de generalización, robustez y adaptabilidad, constituyéndose en una alternativa poderosa a los

modelos más complejos como los basados en redes neuronales.

6.2.4. Modelos secuenciales

Alam et al. (2021) comparan un modelo de aprendizaje profundo diseñado para datos temporales en panel con un modelo clásico de riesgo discreto, mostrando que los enfoques secuenciales pueden lograr niveles superiores de sensibilidad en la predicción de quiebras. Los modelos secuenciales son especialmente adecuados para capturar dinámicas temporales complejas en series financieras. A diferencia de los enfoques tradicionales, estos modelos aprenden representaciones internas que permiten identificar patrones recurrentes o dependencias a lo largo del tiempo.

En esa línea, Nayak y Rout (2023) evalúan el desempeño de redes LSTM, CNN y ANN para la predicción de insolvencias, demostrando que incluso arquitecturas simples como las redes neuronales artificiales pueden superar en F1-score a modelos secuenciales más complejos, siempre que se implementen adecuadamente esquemas de remuestreo y normalización. Aunque sus resultados favorecen la ANN, el uso de trayectorias históricas sigue siendo central en su metodología, lo que refuerza la utilidad de los modelos temporales en esta tarea.

En esta sección se incluyen variantes basadas en redes neuronales como LSTM, BiLSTM y CNN-LSTM, tanto en su forma básica como en versiones enriquecidas con variables categóricas (*dummies*) o estructuras apiladas de ventanas de cinco años.

Tabla 6.4: *Desempeño promedio de modelos secuenciales (validación cruzada 10 folds)*

Modelo	Accuracy	Precision	Recall	F1-score	AUC	LogLoss
LSTM	0,5299	0,1729	0,7187	0,2788	0,6539	0,6665
BiLSTM	0,5872	0,1816	0,6462	0,2836	0,6612	0,6207
BiLSTM + Dummies	0,6931	0,2057	0,4988	0,2912	0,6568	0,5970
CNN-LSTM + Dummies	0,6521	0,2039	0,6032	0,3048	0,6728	0,6217
LSTM (ventana 5 años)	0,7085	0,8630	0,7200	0,7850	0,7569	0,5894
CNN-LSTM (ventana 5 años)	0,6976	0,8925	0,6718	0,7666	0,7931	0,5782

Fuente: Elaboración propia con base en resultados de validación cruzada estratificada (10 folds).

En general, los modelos secuenciales básicos como LSTM y BiLSTM lograron niveles de *recall* aceptables, aunque con *accuracy* y F1-score relativamente bajos. La incorporación de variables categóricas mejoró de forma marginal su

capacidad predictiva.

Las versiones basadas en ventanas de cinco años ofrecieron los mejores desempeños dentro del grupo, con F1-scores superiores al 76 % y AUC cercanos a 0,80. Estos resultados confirman que la inclusión explícita de trayectoria financiera en modelos secuenciales mejora

6.2.5. Modelos híbridos

Alaka et al. (2018) realizan una revisión sistemática de modelos de predicción de quiebra y resaltan la importancia de enfoques híbridos para mejorar el desempeño frente a técnicas individuales. En línea con esta perspectiva, los modelos híbridos integran diversas técnicas de aprendizaje para potenciar la capacidad predictiva, aprovechando las fortalezas de algoritmos complementarios como árboles de decisión, redes neuronales o métodos de remuestreo. Estas arquitecturas pueden adoptar formas diversas, incluyendo esquemas en dos etapas, ensamblajes directos o el uso de metaheurísticas para optimización. Por ejemplo, Ansari et al. (2020) proponen un enfoque basado en metaheurísticas (MOA-PSO) para entrenar redes neuronales, logrando mejoras sustanciales en precisión y eficiencia computacional. De forma similar, du Jardin (2016) desarrollan una estrategia en dos fases que emplea mapas de Kohonen para segmentar empresas, combinados con modelos específicos por grupo y técnicas ensemble como bagging y boosting para incrementar la estabilidad. Asimismo, Zelenkov et al. (2017) plantean un modelo híbrido que utiliza algoritmos genéticos tanto para la selección de variables como para la ponderación interna de clasificadores, alcanzando un buen equilibrio entre sensibilidad y especificidad en un conjunto de firmas rusas. En investigaciones más recientes, M. et al. (2025) combinan XGBoost con redes neuronales artificiales (ANN) sobre datos polacos desequilibrados, logrando mejoras sustanciales en precisión y F1-score. Por su parte, S. Wang y Chi (2024) introducen una arquitectura que fusiona redes adversarias generativas (GAN) con modelos de boosting como LightGBM, optimizada para escenarios con alta desproporción de clases. Finalmente, Amirshahi y Lahmiri (2024) implementan un ensamble optimizado que combina XGBoost, LightGBM y CatBoost, ajustado mediante validación cruzada sobre un conjunto altamente desbalanceado, y demuestran que esta combinación híbrida supera en desempeño a cualquier modelo individual previo, evidenciando la vigencia y potencia de los esquemas híbridos en tareas de predicción de quiebra.

En esta sección se presentan cuatro modelos híbridos: un esquema de *stacking* entre regresión logística, LightGBM y XGBoost (todos con SMOTE), un ensamblaje directo de XGBoost con una red neuronal artificial, una combinación de GAN con LightGBM, y un ensamble reciente propuesto por Amirshahi y Lahmiri (2024), el cual integra XGBoost, LightGBM y CatBoost mediante un clasificador meta de regresión logística.

Tabla 6.5: *Desempeño promedio de modelos híbridos (validación cruzada 10 folds)*

Modelo	Accuracy	Precision	Recall	F1-score	AUC	LogLoss
Stacking (Logit + LGBM + XGB + SMOTE)	0,8334 ± 0,0051	0,3341 ± 0,0197	0,2479 ± 0,0087	0,2845 ± 0,0123	0,6953 ± 0,0113	0,4209 ± 0,0084
XGBoost + ANN	0,8029 ± 0,0181	0,8237 ± 0,0144	0,9345 ± 0,0185	0,8755 ± 0,0114	0,8959 ± 0,0202	1,0073 ± 0,1686
GAN + LightGBM	0,8317 ± 0,0176	0,8841 ± 0,0162	0,8901 ± 0,0199	0,8869 ± 0,0121	0,9003 ± 0,0161	0,3725 ± 0,0451
Ensamble híbrido (XGB + LGBM + CatB)	0,8705 ± 0,0024	0,5635 ± 0,0338	0,1264 ± 0,0277	0,2054 ± 0,0394	0,7586 ± 0,0116	0,3388 ± 0,0052

Fuente: Elaboración propia con base en resultados de validación cruzada estratificada (10 folds).

Los modelos híbridos presentaron un rendimiento destacado, especialmente aquellos que combinan redes neuronales y árboles de decisión. La arquitectura XGBoost + ANN logró un F1-score promedio de 0,8755, mientras que GAN + LightGBM alcanzó el mejor desempeño global dentro del grupo, con un F1-score cercano al 89 % y una AUC de 0,9003. Por su parte, el ensamble híbrido entre XGBoost, LightGBM y CatBoost obtuvo una AUC de 0,7586 y un desempeño competitivo, destacando la integración de múltiples modelos base dentro de una arquitectura apilada como la sugerida por Amirshahi y Lahmiri (2024).

Estas combinaciones sugieren que el uso de ensamblajes y redes profundas puede ser altamente eficaz para tareas de predicción de quiebra, en la medida en que se cuenta con datos estructurados suficientes y técnicas robustas de remuestreo.

6.3. Síntesis comparativa y discusión final

La Tabla 6.6 resume los cinco modelos con mejor desempeño en términos de F1-score, considerando todas las arquitecturas evaluadas. Si bien los modelos como LightGBM y XGBoost entrenados sobre ventanas de cinco años obtuvieron los puntajes más altos (F1-score superiores a 0,89), el modelo funcional propuesto alcanzó un valor competitivo de **0,8497**, con la **recall más alta del grupo** (94,46 %) y una desviación estándar notablemente baja. Esto indica no solo un buen equilibrio entre precisión y sensibilidad, sino también una alta consistencia en los resultados.

Tabla 6.6: *Modelos con mayor desempeño en F1-score para predicción de riesgo empresarial*

Modelo	F1-score	AUC	LogLoss	Recall	Precision	Desv. F1
LightGBM (ventana 5 años)	0.9002	0.9010	0.3476	0.8840	0.9173	±0.0101
XGBoost (ventana 5 años)	0.8954	0.8996	0.3519	0.8824	0.9091	±0.0120
GAN + LightGBM	0.8869	0.9003	0.3725	0.8901	0.8841	±0.0121
XGBoost + ANN	0.8755	0.8959	1.0073	0.9345	0.8237	±0.0114
k-NN funcional optimizado (propuesto)	**0.8497**	**0.8474**	**0.8210**	**0.9446**	**0.7723**	**±0.0081**

En términos de robustez, el modelo funcional mostró estabilidad tanto en la curva de aprendizaje como en el análisis *bootstrap*, manteniendo métricas consistentes incluso con muestras pequeñas y bajo condiciones de alta variabilidad. Esta propiedad lo posiciona como una herramienta confiable en entornos donde el volumen de datos es limitado o donde se requiere adaptabilidad a diferentes contextos.

Una de las diferencias más relevantes entre el modelo funcional propuesto y las arquitecturas avanzadas evaluadas es su capacidad de interpretación nativa. Mientras que modelos como XGBoost, LightGBM o redes neuronales requieren herramientas externas como SHAP o LIME para estimar la importancia de las variables, el enfoque funcional permite interpretar directamente las decisiones del clasificador a partir de la estructura misma de la métrica utilizada.

La métrica funcional diseñada en esta tesis asigna pesos diferenciados a cada indicador financiero, lo que permite descomponer la distancia entre empresas en términos de sus componentes contables. Este mecanismo no solo mejora el ajuste del modelo, sino que habilita una lectura directa de cuáles son las variables que más influyen en la similitud entre trayectorias. Así, el modelo no se limita a emitir una predicción binaria, sino que también permite

entender en qué dimensiones se asemejan dos empresas y por qué se clasifica a una como en riesgo.

Además, el uso del algoritmo k-NN aporta una segunda capa de interpretabilidad: es posible identificar explícitamente a qué empresas del pasado se parece cada observación actual, y basar la decisión en ejemplos reales observados. Esta característica de trazabilidad por similitud refuerza la transparencia del modelo y puede ser clave en escenarios de supervisión financiera, auditoría o asesoría contable, donde es necesario justificar y contextualizar cada predicción.

En conjunto, aunque el modelo funcional no lidera el ranking por F1-score, sí ofrece un equilibrio único entre rendimiento, robustez, interpretabilidad y trazabilidad. Esta combinación lo convierte en una opción valiosa para sistemas de alerta temprana, monitoreo financiero sectorial y escenarios donde el entendimiento del fenómeno es tan importante como su predicción.

6.4. Reproducibilidad de los modelos comparativos

Con el fin de garantizar la trazabilidad de los resultados y permitir la replicación completa del análisis comparativo, todos los modelos tradicionales, dinámicos y avanzados evaluados en este capítulo han sido implementados en Python y organizados en un conjunto de scripts disponibles públicamente.

Estos scripts replican paso a paso los procesos de limpieza, partición, validación cruzada, entrenamiento y evaluación de cada modelo. Se emplearon técnicas consistentes en todo el análisis, incluyendo el uso controlado de SMOTE para desbalance, curvas ROC y PRC promedio, métricas estándar de clasificación y reportes automatizados para facilitar la comparación entre enfoques.

- **Repositorio GitHub:** https://github.com/tesisluisruizparedes/Tesis_KNN_Funcional
- **Carpeta `scripts_comparativos/`:** implementación de modelos tradicionales (Logit, Probit, k-NN), modelos dinámicos (ventanas móviles) y modelos avanzados (XGBoost, Random Forest, híbridos con redes neuronales)
- **Entrada común:** base con variables reducidas e indicadores financieros (`9_Base_Modelos_Predictivos_Turismo_Reducida_Completa.parquet`)

El código ha sido estructurado para permitir la modificación sencilla de parámetros, inclusión de nuevos modelos y replicación de todos los experimentos presentados. De esta manera, se busca facilitar futuras investigaciones que deseen extender el análisis a otros sectores económicos, países o variables financieras.

Capítulo 7

Conclusiones

Esta investigación propuso y validó una métrica funcional multivariada personalizada, integrada en un clasificador *k-Nearest Neighbors* (k-NN) funcional, para predecir el riesgo de quiebra empresarial. Se implementaron dos versiones del modelo: una versión base, construida exclusivamente sobre trayectorias financieras, y una versión extendida que incorpora variables estructurales (sector económico, región y desfase temporal) directamente en la métrica.

En la versión base, con parámetros óptimos ($k = 15$, $\lambda \approx 1.99$), el modelo alcanzó un $F1$-score promedio de 0,8497 y un *recall* de 94,46 %, minimizando falsos negativos y garantizando una alta sensibilidad, crítica en sistemas de alerta temprana. Las alertas emitidas correspondieron a casos reales de deterioro financiero en un 77 %, como lo muestra la precisión, con baja variabilidad entre validaciones cruzadas. Estos resultados posicionan al modelo como un complemento valioso en contextos donde se priorizan la interpretabilidad, la trazabilidad de decisiones o la disponibilidad limitada de datos, aspectos que la literatura reciente ha señalado como esenciales para la adopción efectiva de modelos predictivos en entornos reales (P. Carmona et al., 2022).

En la versión extendida, al incorporar atributos cualitativos y temporales en la métrica funcional, el modelo alcanzó un desempeño aún más alto, con un *F1*-score de **0,9802**, superando ampliamente los enfoques tradicionales y acercándose a los mejores resultados reportados en la literatura nacional, como el estudio de Correa (2023), que también utiliza variables categóricas y técnicas avanzadas sobre datos financieros colombianos. Aunque los contextos y variables no son exactamente equivalentes, esta coincidencia refuerza la solidez del enfoque propuesto y su potencial para escenarios complejos.

Al comparar con otras métricas funcionales (como L_1, L_2, Mahalanobis o PCA) reveló una caída en el rendimiento predictivo, lo que valida la superioridad de la distancia diseñada. Esto refuerza el carácter diferencial de la métrica desarrollada, cuyas propiedades fueron concebidas para operar de manera conjunta con un espacio funcional diseñado específicamente para este fin, permitiendo capturar similitudes económicamente relevantes que otras distancias ignoran.

Las métricas estándar no fueron aplicadas sobre el espacio funcional $\mathcal{F}$, ya que este fue construido de manera conjunta con la métrica propuesta, conformando una estructura indivisible. Aplicar métricas sobre $\mathcal{F}$ sería metodológicamente incorrecto, pues implicaría ignorar los supuestos estructurales bajo los cuales dicho espacio fue concebido, entre ellos la tolerancia explícita a trayectorias incompletas y valores infinitos —condiciones que las métricas convencionales no pueden procesar adecuadamente—, así como una estructura explícita de dependencia temporal movilizada. Por esta razón, las métricas convencionales se evaluaron exclusivamente sobre trayectorias discretas homogéneas, definidas en espacios vectoriales clásicos equivalentes a $\mathbb{R}^n$, sin incorporar alineación temporal ni continuidad relativa en las observaciones, garantizando así la validez metodológica y la coherencia interna de la comparación empírica.

Más allá de utilizar un algoritmo clásico como el k-NN, esta tesis propone una métrica funcional multivariada personalizada, desarrollada desde cero, que incorpora acumulación tipo L_1, penalización por pérdida de dominio, acotamiento racional de extremos y combinación ponderada por indicador. Esta construcción redefine tanto el espacio funcional como el criterio de similitud, transformando al k-NN en una herramienta trazable e interpretable, y superando una de sus limitaciones históricas.

Además, mostró alta estabilidad ante cambios muestrales: tanto en validación cruzada como en pruebas *bootstrap*, la variación en las principales métricas de desempeño (como el F_1-score, la precisión y el *recall*) fue mínima, y el modelo mantuvo eficacia incluso con subconjuntos reducidos, como lo confirman las curvas de aprendizaje.

El modelo también se destaca por su interpretabilidad. A diferencia de modelos de caja negra, el k-NN funcional permite descomponer las decisiones en dos niveles: (1) por importancia relativa de indicadores —identificada mediante la optimización de pesos— y (2) por similitud con ejemplos reales de empresas históricamente comparables. Esta trazabilidad se completa al identificar explícitamente los vecinos más cercanos a cada empresa evaluada,

lo que permite explicar cada predicción mediante analogías comprensibles para analistas y directivos.

Se realizó una limpieza estructural profunda, incluyendo la depuración y trazabilidad del NIT, la estandarización geográfica y sectorial, y la imputación contable basada en reglas verificables. Además, se aplicó una reducción de dimensionalidad mediante agrupamiento jerárquico y VIF, seleccionando 17 indicadores representativos de los 45 originales, manteniendo la diversidad temática y favoreciendo la interpretabilidad sin sacrificar desempeño. Esta estrategia metodológica busca también responder a una de las debilidades más frecuentes en la literatura sobre predicción de quiebra en PYMEs, como señalan Cheraghali y Molnár (2024), quienes evidencian que más del 37 % de los estudios no emplean técnicas adecuadas de selección de variables.

Para garantizar la comparabilidad a lo largo del tiempo, se integraron de forma coherente los reportes financieros de empresas colombianas entre 1995 y 2023, todos provenientes de la Superintendencia de Sociedades. Aunque durante este periodo se presentaron cambios en las normas contables, se aplicaron procesos de normalización técnica que permitieron conservar la continuidad conceptual de los indicadores financieros utilizados en el modelo.

Finalmente, el estudio garantizó la reproducibilidad total de sus resultados. Todos los scripts fueron documentados y publicados en un repositorio GitHub abierto, y se desarrolló una aplicación web interactiva mediante *Streamlit* para aplicar el modelo sobre nuevos datos. Estas acciones no solo fortalecen la transparencia y trazabilidad del enfoque, sino que también facilitan su adopción práctica por parte de empresas, reguladores o investigadores.

7.1. Cumplimiento de los objetivos

El desarrollo de esta investigación permitió alcanzar de manera verificable tanto el objetivo general como los objetivos específicos definidos al inicio del proyecto. En conjunto, estos logros consolidan la validez del enfoque propuesto y responden de forma directa a la pregunta de investigación.

En primer lugar, se cumplió el objetivo general: desarrollar y validar un enfoque metodológico para la predicción del riesgo de quiebra en empresas del sector turismo colombiano, basado en trayectorias financieras multivariadas e integrado en un modelo k-NN funcional con métrica personalizada. El modelo alcanzó niveles de precisión comparables a los de técnicas avanzadas como XGBoost y LSTM, ofreciendo además ventajas clave en transparencia, trazabilidad y facilidad de implementación.

El primer objetivo específico —diseñar una métrica funcional multivariada que cuantificara la similitud entre trayectorias financieras— se cumplió en dos fases. En una primera versión, la métrica fue definida como una distancia tipo L_1 acumulada, con penalización por pérdida de dominio, acotamiento racional para mitigar *outliers* y pesos optimizados por indicador. Posteriormente, se extendió esta métrica para incorporar variables cualitativas y temporales (sector económico, región geográfica y desfase de tiempo), lo que mejoró sustancialmente el desempeño predictivo sin sacrificar interpretabilidad. Esta evolución metodológica demuestra la flexibilidad del enfoque propuesto y su capacidad de adaptarse a estructuras de datos más complejas.

En segundo lugar, se construyó un conjunto de datos robusto a partir de los estados financieros reportados por empresas turísticas ante la Superintendencia de Sociedades entre 1995 y 2023. Se calcularon 45 indicadores, y mediante agrupamiento jerárquico y VIF se seleccionaron 17 variables representativas. La variable dependiente `RQ_final` se definió como proxy del riesgo de quiebra, combinando criterios institucionales y evidencia de cierre operativo.

El tercer objetivo —evaluar el desempeño predictivo del modelo funcional y compararlo con enfoques alternativos— se cumplió mediante un protocolo riguroso de validación cruzada estratificada de 10 pliegues, aplicado tanto a la versión base como a la extendida. En todos los casos se usó la misma base reducida de 17 indicadores para garantizar control metodológico. El modelo funcional superó a los modelos tradicionales (como regresión logística y Probit), y mostró un rendimiento competitivo frente a técnicas más sofisticadas como Random Forest, XGBoost y LSTM. Esta superioridad no solo se reflejó en el *F1*-score promedio, sino también en la baja variabilidad entre

folds y en su estabilidad frente a diferentes tamaños muestrales. Además, la versión extendida del modelo, al incorporar atributos estructurales, logró un desempeño aún más alto (*F1*-score de **0,9802**), acercándose a los mejores resultados reportados en la literatura nacional (Correa, 2023).

Finalmente, el cuarto objetivo fue desarrollar una herramienta computacional que facilitara la aplicación práctica del modelo. Se implementó una aplicación web interactiva con `Streamlit`, que permite ingresar trayectorias contables, visualizar vecinos más cercanos y observar las variables más influyentes en la predicción. Esta herramienta fue publicada en un repositorio abierto para facilitar su replicabilidad y escalabilidad.

7.2. Aportes al conocimiento y a la práctica

Como respuesta al desafío de construir modelos predictivos que sean, al mismo tiempo, efectivos e interpretables, esta investigación desarrolla —desde una perspectiva teórico–metodológica— un enfoque basado en la construcción de un espacio funcional compuesto por trayectorias financieras. Sobre este espacio se define una métrica personalizada que permite evaluar la similitud entre empresas considerando la evolución completa de sus indicadores. A partir de esta noción de distancia, se implementó un clasificador funcional del tipo k-NN, que integra el espacio $\mathcal{F}$, la métrica y el algoritmo. Esta arquitectura permitió alcanzar niveles de rendimiento comparables con métodos avanzados de predicción del riesgo, manteniendo al mismo tiempo una estructura interpretativa y transparente. Este diseño representa, además, una ventaja metodológica concreta, al combinar principios del análisis funcional con una técnica no paramétrica.

Un primer componente clave de esta arquitectura fue la construcción del espacio funcional $\mathcal{F}$, utilizando ventanas móviles retrospectivas de longitud fija. Esta estrategia permitió uniformar la representación temporal de las trayectorias financieras sin recurrir a procesos extensivos de imputación. Mientras que trabajos recientes suelen emplear dichas ventanas como insumos directos para modelos supervisados tradicionales Abrahamsen et al. (2024), en este caso se incorporaron como parte de una estructura funcional multivariada, diseñada para facilitar la comparación de patrones a lo largo del tiempo. Este enfoque ofrece una representación coherente y comparativa que se adapta bien a las particularidades de los registros contables reales, y aporta una perspectiva distinta dentro del análisis funcional aplicado a finanzas.

Esta métrica incorpora cuatro elementos clave —acumulación de diferencias a lo largo del tiempo, penalización por pérdida de dominio, acotamiento de extremos y combinación ponderada de indicadores— que permiten evaluar la similitud entre empresas sin necesidad de imputación forzada ni pérdida de escala original. Aunque formalmente se trata de una semi-métrica, su formulación sobre un marco funcional continuo garantiza robustez, interpretabilidad y escalabilidad a contextos con mayor granularidad temporal (mensual o diaria) sin necesidad de rediseños estructurales.

Al integrarse en un clasificador funcional, esta construcción transforma al k-NN en una herramienta robusta, sensible e interpretable. A diferencia de distancias estándar (L1, L2, Mahalanobis, derivadas, PCA), la métrica propuesta logra un desempeño predictivo superior en condiciones controladas,

como se demuestra en los experimentos presentados en esta tesis. Este diseño no solo contribuye a la literatura sobre predicción de riesgo, sino que representa una innovación metodológica en el campo del análisis financiero funcional.

Aunque en esta tesis fue implementada dentro de un clasificador k-NN, la métrica desarrollada constituye en sí misma un aporte metodológico independiente. Su diseño permite comparar trayectorias financieras completas entre empresas, respetando su forma y evolución temporal, incluso en presencia de datos faltantes o valores extremos. Esta capacidad abre la puerta a su uso en otras tareas que también se basan en la noción de similitud, como el agrupamiento de empresas con trayectorias similares, la detección de trayectorias atípicas o la caracterización de trayectorias representativas dentro de un conjunto. si bien no cumple estrictamente con la desigualdad triangular, su diseño responde a los principios de las semi-métricas generalizadas frecuentemente utilizadas en el análisis funcional moderno, lo que valida su aplicabilidad en tareas de clasificación y clustering (N. James et al., 2023).

En términos empíricos, los resultados obtenidos validan la solidez y trazabilidad del modelo propuesto frente a alternativas existentes. Una de las contribuciones más destacadas de este trabajo reside en el fortalecimiento de la interpretabilidad del clasificador funcional k-NN. A diferencia de su versión tradicional, que no permite evaluar la importancia relativa de cada variable, el modelo propuesto incorpora una métrica personalizada con pesos diferenciados por indicador financiero, los cuales fueron optimizados durante el proceso de validación. Esta estructura permite descomponer cada predicción tanto por variable —identificando, por ejemplo, los indicadores con mayor peso específico según la versión evaluada del modelo— como por ejemplos análogos, es decir, empresas con trayectorias históricas similares. Esta doble dimensión de trazabilidad facilita la explicación de los resultados incluso para usuarios no técnicos y elimina la necesidad de técnicas externas como SHAP. Para reforzar esta ventaja, se incluyó un ejemplo detallado que muestra el funcionamiento paso a paso de la métrica, tanto con trayectorias sintéticas como con un caso real, lo que facilita su comprensión y valida su aplicabilidad en contextos contables.

Más allá del alto desempeño sobre el conjunto de datos, el modelo también demostró una notable estabilidad ante distintas condiciones muestrales, como se evidenció mediante validación cruzada y *bootstrap*. La curva de aprendizaje sugiere que el modelo es robusto, pero posiblemente sobreajusta con muestras pequeñas, y que la métrica funcional mantiene un desempeño razonablemente estable incluso cuando se incrementa el tamaño de la muestra y disminuye la

variabilidad. Esta solidez operativa le confiere una ventaja frente a enfoques más complejos, como las redes neuronales profundas, que suelen requerir grandes volúmenes de datos, calibración intensiva y son más sensibles a los cambios en la composición del conjunto de entrenamiento.

El estudio proporciona evidencia rigurosa sobre la aplicabilidad del modelo propuesto en el contexto colombiano, específicamente en el sector turismo. Si bien se trata de un caso sectorial, su análisis muestra que el enfoque puede implementarse con éxito en contextos contables reales, aportando una alternativa interpretativa y replicable.

En términos prácticos, esta tesis también ofrece aportes concretos tanto para la práctica profesional como para la comunidad académica, mediante herramientas funcionales y documentación reproducible. Esto no solo fomenta la transparencia científica, sino que facilita la adaptación de la metodología a otros sectores o países. Además, se desarrolló una metodología robusta para la construcción de bases contables depuradas, que incluye estrategias de imputación financieramente coherentes, control de valores extremos y un flujo de trabajo sistemático que abarca ventanas móviles, validación cruzada estratificada, reducción estructurada de variables y optimización bayesiana con `Optuna`. Estos elementos aseguran que tanto los resultados como el proceso puedan ser auditados, replicados y aprovechados por otros equipos bajo estándares rigurosos de calidad.

Por último, se entrega una base sectorial limpia, estructurada y validada para el sector turismo en Colombia, con potencial de reutilización en tareas de vigilancia financiera, estudios comparativos o formulación de política sectorial.

En cuanto a la práctica profesional, el modelo desarrollado puede ser utilizado como una herramienta de alerta temprana accesible, explicable y de bajo costo computacional. Su implementación en una aplicación interactiva mediante *Streamlit* demuestra su potencial de adopción por parte de entidades con capacidades técnicas limitadas. Si bien requiere una calibración inicial —particularmente en los pesos asociados a los indicadores financieros—, una vez realizado este ajuste, el modelo puede incorporar nuevos datos sin necesidad de ser reentrenado. Esta propiedad resulta especialmente útil para entornos que demandan monitoreo continuo. A diferencia de modelos avanzados como XGBoost o LSTM, cuyo mantenimiento exige recursos computacionales elevados y conocimientos especializados, el enfoque propuesto puede ser implementado y actualizado sin dichas barreras, facilitando su integración en prácticas reales de gestión financiera.

El verdadero aporte metodológico de esta tesis no radica en el uso del

algoritmo k-NN como tal, sino en la construcción de una métrica funcional multivariada personalizada, desarrollada desde cero, que redefine tanto el espacio de representación como el criterio de similitud entre empresas. Esta transformación metodológica permitió que un modelo clásico, usualmente limitado por su simplicidad, alcanzara niveles de desempeño comparables con métodos avanzados, sin renunciar a la interpretabilidad.

Estos aportes confirman que la tesis no solo genera nuevo conocimiento teórico y metodológico, sino que también ofrece soluciones prácticas, replicables y adaptables a problemas reales.

7.2.1. Comparación con la literatura reciente

La literatura más reciente en predicción de quiebra utiliza principalmente métodos de *machine learning* avanzados como técnicas de *boosting*, ensamblados y redes neuronales profundas. Por ejemplo, Amirshahi y Lahmiri (2024) analizan modelos de tipo *gradient boosting* (XGBoost, LightGBM y CatBoost), y demuestran que un ensamblaje óptimo entre ellos supera consistentemente a sus versiones individuales, particularmente en métricas como AUC y precisión. Este tipo de arquitectura fue replicada también en esta investigación, sobre la base de empresas turísticas, lo que permitió una comparación directa.

De manera similar, Papík et al. (2022) aplican el algoritmo CatBoost sobre una amplia muestra de PYMEs, logrando un AUC de 0.9812 al combinar variables categóricas con razones financieras. Sus resultados muestran que la inclusión de información cualitativa mejora sustancialmente la capacidad predictiva frente a modelos basados únicamente en indicadores numéricos. Aunque en esta tesis no se incorporaron variables categóricas externas al ámbito financiero, se replicaron modelos basados en árboles de decisión como LightGBM, XGBoost y Catboost sobre una base estructurada de indicadores contables, lo cual permite contrastar la capacidad de estos algoritmos en contextos distintos y evaluar su robustez en dominios más acotados.

De igual modo, Abrahamsen et al. (2024) emplean LightGBM sobre datos financieros trimestrales de empresas nórdicas, utilizando una estructura de ventana móvil, y logran la mayor precisión comparada con otros métodos clásicos y de *machine learning*. Los autores destacan la capacidad de los modelos de *boosting* para capturar dinámicas temporales complejas y adaptarse a distintas condiciones macroeconómicas.

En general, los estudios recientes que emplean algoritmos de ensamblado basados en árboles —como *Random Forest*, XGBoost, LightGBM o CatBoost— reportan desempeños predictivos muy elevados, con métricas de precisión y AUC cercanas al 95–98 % (Abrahamsen et al., 2024; Amirshahi & Lahmiri, 2024; Papík et al., 2022). Sin embargo, estos modelos suelen depender de una cuidadosa ingeniería de características, una selección exhaustiva de hiperparámetros y, en muchos casos, estrategias de balanceo adaptadas al conjunto de entrenamiento. En contraste, el modelo funcional propuesto en esta tesis adopta una aproximación radicalmente distinta: trabaja directamente con trayectorias financieras multivariadas, sin necesidad de descomposición o selección de variables previa, y aplica un clasificador no paramétrico tipo k-NN cuya potencia predictiva proviene de una métrica especializada de similitud

temporal. Pese a su mayor sencillez estructural y ausencia de ensamblados complejos, este enfoque logra desempeños competitivos, lo que refuerza su viabilidad en contextos con recursos computacionales limitados o necesidades de interpretabilidad clara.

Por otro lado, varios estudios recientes han recurrido a redes neuronales profundas para abordar el problema de predicción de quiebras empresariales. Nayak y Rout (2023) comparan el desempeño de tres arquitecturas distintas —redes neuronales artificiales (ANN), redes convolucionales (CNN) y redes de memoria a largo plazo (LSTM)— sobre un conjunto de datos históricos de empresas estadounidenses (1971–2017). Los autores aplican distintas técnicas de balanceo de clases y encuentran que la ANN obtiene sistemáticamente mejores resultados que las otras dos arquitecturas, especialmente cuando se combina con estrategias híbridas de muestreo como *SMOTE+Tomek links*. Estos hallazgos destacan el impacto del preprocesamiento y del tipo de red seleccionada en la precisión del modelo, así como las limitaciones de enfoques puramente secuenciales como LSTM en contextos contables tradicionales. En contraste, el modelo funcional propuesto en esta tesis no requiere redes neuronales profundas ni procedimientos complejos de entrenamiento, y aun así logra desempeños competitivos, especialmente en *recall*, sin sacrificar interpretabilidad.

Song et al. (2025) proponen un modelo híbrido que combina redes neuronales convolucionales (CNN) con una arquitectura bidireccional LSTM y un mecanismo de atención, optimizado mediante *Hyperband*, y aplicado sobre datos financieros de empresas chinas con técnicas de remuestreo tipo SMOTE. El modelo alcanza una precisión del 99,4 % en la predicción de *financial distress*, lo que evidencia el potencial de las arquitecturas profundas híbridas cuando se ajustan adecuadamente sobre conjuntos temporales complejos.

De igual modo, H. Zhang y Zhang (2025) integran XGBoost con redes neuronales convolucionales (CNN) y una arquitectura BiLSTM en un solo *framework* aplicado a datos del índice S&P 500, logrando métricas de desempeño notablemente altas, con una precisión aproximada del 93,8 % y un *recall* cercano al 95,8 %. Este enfoque destaca por su capacidad de combinar mecanismos de representación jerárquica con secuencias temporales y aprendizaje basado en árboles.

Estos estudios confirman que las redes profundas modernas (LSTM, CNN) son altamente efectivas en la predicción de quiebra y que al combinarse con algoritmos de tipo árbol, pueden mejorar se desempeñó. Sin embargo, este rendimiento suele estar condicionado a arquitecturas complejas y baja traza-

bilidad en sus decisiones internas. En comparación, nuestro enfoque funcional multivariado ofrece un avance conceptual distinto: representa cada empresa como una función financiera continua en el tiempo, lo que permite comparar trayectorias completas con interpretabilidad explícita a nivel de indicador. Esta formulación facilita tanto el análisis como la validación posterior de los resultados, otorgando mayor transparencia y control sobre la lógica del modelo predictivo.

Los trabajos híbridos y de ensamblado también proliferan. Muslim y Yosza (2021) usan un modelo de *stacking* cuyos modelos base incluyen k-NN, SVM, árboles de decisión (DT) y *random forest* (RF), con LightGBM como meta-aprendiz, alcanzando una precisión cercana al 97 %, superior a cada modelo base individual. De manera similar, M. et al. (2025) aplican en conjunto XGBoost y una red neuronal artificial (ANN) sobre datos polacos desbalanceados; su modelo híbrido XGBoost–ANN obtiene los mejores resultados globales en precisión y *F1-score* frente a clasificadores tradicionales.

Recientes propuestas van más allá e incorporan *GANs* para generar datos sintéticos minoritarios. Por ejemplo, S. Wang y Chi (2024) combinan redes adversarias generativas con técnicas de *oversampling* y un sistema de *stacking*, logrando un AUC cercano al 90 % en tareas de *credit scoring*.

Estos enfoques demuestran que estrategias de *ensemble* avanzadas (como *stacking*, fusión de *boosting* con redes neuronales, o *GANs* combinadas con LightGBM) pueden mejorar sustancialmente la predicción de quiebra empresarial. Sin embargo, a diferencia de estos sistemas complejos, el k-NN funcional multivariado propuesto en esta tesis constituye un método no paramétrico que se beneficia directamente de la estructura funcional inherente a los datos. Conceptualmente, esto representa un avance metodológico relevante, ya que extiende la distancia clásica del k-NN al dominio funcional multivariado, capturando tendencias temporales sin requerir grandes volúmenes de entrenamiento ni arquitecturas sofisticadas.

La evidencia reciente coincide en que los modelos de aprendizaje automático avanzados superan ampliamente a los enfoques tradicionales en la predicción de quiebra empresarial. Por ejemplo, Nayak y Rout (2023) subrayan el auge de arquitecturas como LSTM y CNN por su capacidad para modelar dependencias temporales complejas, mientras que otros hallazgos destacan que combinaciones de técnicas como XGBoost+CNN+BiLSTM alcanzan niveles de precisión muy elevados (Song et al., 2025).

Los resultados obtenidos en esta investigación muestran que el modelo funcional k-NN optimizado logra un desempeño comparable frente a dichos

modelos avanzados, lo que posiciona al enfoque funcional multivariado como un aporte relevante en términos aplicados, especialmente considerando que muchos de estos modelos fueron implementados sobre los mismos datos, permitiendo una comparación directa bajo condiciones homogéneas. En particular, el desempeño alcanzado por el modelo funcional extendido (F1-score de 0,9802) es equivalente al reportado por Correa (2023), quien aplica XGBoost sobre empresas colombianas, incorporando también variables estructurales como el tamaño, el sector y la región. Esta cercanía en resultados valida empíricamente la solidez del enfoque propuesto, con la ventaja adicional de ofrecer mayor interpretabilidad y menor dependencia de ingeniería de características.

Así, frente a la literatura reciente que explora ensambles de árboles de decisión, redes neuronales recurrentes o convolucionales, modelos híbridos y técnicas como el remuestreo o el *stacking*, la metrica propuesta le permitio al modelo funcional k-nn alcanzar rendimientos comparables con métodos avanzados de predicción, manteniendo al mismo tiempo una estructura interpretativa y transparente, ofreciendo un aporte significativo a la predicción temprana del riesgo empresarial.

7.3. Limitaciones y proyecciones futuras

A pesar de los sólidos resultados obtenidos, esta investigación reconoce una serie de limitaciones metodológicas y contextuales que delimitan el alcance del modelo propuesto. Lejos de comprometer la validez de los hallazgos, estas limitaciones reflejan la complejidad del problema abordado y permiten delinear rutas claras para futuras investigaciones que amplíen, adapten o refinen el enfoque planteado.

Una limitación inherente al enfoque propuesto es que el clasificador k-NN funcional no genera probabilidades calibradas ni modela relaciones no lineales entre los indicadores financieros. La clasificación se basa únicamente en la proporción de vecinos con historial de quiebra y en la similitud entre trayectorias evaluada por indicador, sin capturar interacciones complejas entre variables. Aunque estas características limitan ciertos usos avanzados —como el cálculo de pérdidas esperadas o la identificación de combinaciones de riesgo—, también forman parte de las fortalezas del modelo, al priorizar la interpretabilidad, la simplicidad y la trazabilidad directa de cada predicción.

Desde el punto de vista temporal, el uso de ventanas móviles retrospectivas ancladas al último año con datos disponibles por empresa permitió representar trayectorias completas, incluyendo aquellas de empresas que ya no estaban activas. Esta estrategia fue deliberadamente escogida, dado que el objetivo principal del estudio es identificar la evolución dinámica de los indicadores antes de la quiebra, por lo que resultaba prioritario capturar la historia reciente de las empresas que efectivamente quebraron, más que mantener una alineación estricta por año calendario. No obstante, esta decisión implica dejar por fuera factores externos con una dimensión temporal definida, como recesiones económicas, reformas fiscales o crisis sectoriales, que afectan simultáneamente a múltiples empresas durante un mismo periodo. Una posible línea futura sería desarrollar métricas híbridas que integren sensibilidad al contexto temporal sin sacrificar la estructura funcional, o bien incorporar variables contextuales (como inflación, tasas de interés o PIB sectorial) como componentes externos adicionales al modelo.

Una posibilidad metodológica que fue abordada parcialmente en esta tesis consiste en la integración de atributos categóricos o constantes dentro de la métrica funcional. En su versión extendida, el modelo incorporó variables como el sector económico (CIIU), el departamento (DEP) y el desfase de la ventana móvil, asignando un peso optimizado a cada componente y ajustando su contribución a la distancia total entre empresas. Esta inclusión demostró

ser efectiva para mejorar el rendimiento predictivo, y permitió ampliar el modelo hacia una estructura funcional-categórica más rica.

No obstante, esta primera integración se implementó mediante funciones de disimilitud discretas estándar (tipo 0/1), sin una adaptación específica a la estructura o semántica subyacente de los datos. Por ejemplo, se consideró como completamente disímil cualquier diferencia entre departamentos, sin tener en cuenta su cercanía geográfica, nivel de urbanización o vinculación turística. De manera similar, el año fue tratado con una distancia lineal simple, sin considerar ciclos económicos ni efectos temporales no lineales.

Como línea futura, se propone avanzar hacia la construcción de métricas categóricas especializadas, análogas a la métrica funcional diseñada en esta tesis para trayectorias contables. Existen múltiples alternativas en la literatura que podrían adaptarse a este propósito. Por ejemplo, se podrían utilizar distancias basadas en jerarquías sectoriales, métricas probabilísticas sensibles a la frecuencia empírica de cada categoría, o distancias semánticas para estructuras categóricas con significado contextual. La distancia de Gower, las medidas basadas en información mutua, o las distancias de árbol aplicadas a clasificaciones como la CIIU representan opciones prometedoras.

Además, sería posible incorporar variables categóricas dinámicas —aquellas que cambian con el tiempo— para capturar eventos como cambios de sede, reorganizaciones o modificaciones en el objeto social de la empresa. En estos casos, el desarrollo de funciones de disimilitud temporales o simbólicas podría enriquecer aún más el modelo.

Este tipo de integración ha sido explorado con éxito en trabajos recientes, como en la métrica EDMD basada en entropía propuesta por Kar et al. (2024), la distancia de Gower modificada de P. Liu et al. (2024), o los modelos funcionales mixtos aplicados a trayectorias financieras y atributos categóricos analizados por Thompson y Davison (2024). Estas referencias muestran que el diseño cuidadoso de las métricas categóricas, lejos de ser un detalle menor, puede marcar una diferencia sustantiva en el rendimiento y aplicabilidad de modelos funcionales en contextos reales.

El enfoque propuesto se limita geográfica y sectorialmente al caso colombiano en el sector turismo. Si bien esto permitió una caracterización profunda y focalizada, restringe la generalización de los hallazgos a otros sectores o países. Futuras investigaciones podrían adaptar el modelo a contextos sectoriales distintos, ajustando los indicadores financieros utilizados, los parámetros óptimos de la métrica, y evaluando su comportamiento bajo otras normativas contables o realidades económicas.

Asimismo, esta tesis no abordó modelos de supervivencia ni arquitecturas especializadas en predicción temporal como Transformers o esquemas *encoder–decoder*, dado que su lógica de salida —orientada a estimar el tiempo restante hasta la quiebra o a generar series completas— no es directamente comparable con el enfoque binario inmediato adoptado aquí. No obstante, estos métodos han demostrado ser efectivos para incorporar explícitamente el horizonte temporal del evento de insolvencia. Por ejemplo, Vallarino (2024) utilizan modelos de Cox funcional para estimar el tiempo hasta la quiebra, mientras que Kuizinienė et al. (2022) destacan la creciente aplicación de arquitecturas tipo Transformer y codificadores secuenciales en el análisis del deterioro financiero.

Como línea futura, se podría explorar la extensión del modelo k-NN funcional hacia entornos temporales más complejos, mediante adaptaciones orientadas a datos censurados o de supervivencia. En este sentido, Guenani et al. (2023) proponen un estimador k-NN robusto para datos ergódicos y sugieren explícitamente su aplicabilidad futura en análisis de supervivencia, lo que abre la posibilidad de desarrollar variantes funcionales capaces de estimar el tiempo restante hasta la quiebra o la evolución del riesgo a lo largo del tiempo.

Desde el punto de vista técnico, aún existen oportunidades de mejora en el diseño y calibración de la métrica. Se podrían explorar nuevas transformaciones de trayectoria (como descomposición en bases de Fourier o wavelets), formas alternativas de penalización por pérdida de dominio, o incluso aplicar técnicas de AutoML y algoritmos evolutivos para optimizar de forma más eficiente los parámetros del modelo.

Cabe señalar que, al aplicar penalización proporcional por pérdida de dominio, la métrica propuesta no cumple estrictamente la desigualdad triangular, por lo que debe entenderse como una semi-métrica penalizada. Esta limitación no invalida su uso en tareas prácticas de clasificación, pero sí sugiere una línea futura orientada al diseño de variantes más formales —sin penalización o con ajustes suaves— que preserven propiedades métricas puras.

En síntesis, cada una de las limitaciones identificadas en esta tesis abre una posibilidad concreta de mejora o ampliación. Este trabajo no constituye un punto de llegada, sino una base metodológica sólida desde la cual avanzar hacia modelos de predicción más robustos, versátiles e interpretables, capaces de responder a desafíos reales en entornos cambiantes y con restricciones prácticas.

Cabe enfatizar que la métrica funcional propuesta en esta tesis fue diseñada

específicamente para capturar la similitud entre trayectorias financieras multivariadas, en el contexto particular de empresas que evolucionan temporalmente hacia estados de deterioro o estabilidad. Su estructura, sus componentes penalizados y la lógica de acumulación están estrechamente ligados a la naturaleza escalonada, discontinua y heterogénea de los estados financieros empresariales. Por esta razón, no sería apropiado aplicar esta misma métrica de forma directa en dominios conceptualmente distintos o en trayectorias construidas a partir de variables cualitativamente diferentes.

Sin embargo, en el campo de la administración existen otros fenómenos donde también se registran trayectorias temporales susceptibles de análisis funcional. Por ejemplo, se pueden construir curvas de evolución en la ejecución presupuestal de entidades públicas, el desempeño operativo de sucursales regionales, los indicadores de calidad en hospitales o universidades a lo largo del tiempo, o los registros mensuales de cumplimiento de metas en proyectos estratégicos. En todos estos casos, sería posible diseñar métricas específicas que capturen la similitud o divergencia en los patrones evolutivos, incorporando penalizaciones, pesos o transformaciones adaptadas al tipo de indicador y al ritmo de actualización de la información.

De esta manera, la contribución metodológica de esta tesis no se limita al problema de predicción de quiebra, sino que ofrece una ruta generalizable para construir modelos funcionales en problemas administrativos donde la evolución temporal multivariada sea central y donde no existan métricas predefinidas adecuadas al contexto.

7.4. Conclusion General

Esta investigación doctoral demuestra que es posible mejorar significativamente la predicción del riesgo de quiebra empresarial mediante una métrica funcional multivariada diseñada específicamente para ese fin. El aporte principal no reside en el uso general de métodos funcionales, sino en la construcción de una función de distancia propia que permite comparar trayectorias contables completas de forma robusta, interpretable y empíricamente eficaz. Esta métrica integra, de manera original, acumulación temporal, penalización por pérdida de dominio, acotamiento de valores extremos y combinación ponderada de indicadores, lo que le permite capturar similitudes económicamente relevantes incluso en presencia de datos faltantes o escalas heterogéneas. Frente a métricas estándar como L1, L2 o Mahalanobis, la propuesta se valida como una alternativa más representativa y adecuada para el análisis de deterioro financiero progresivo.

Uno de los hallazgos más importantes fue constatar que esta métrica, aplicada dentro de un modelo k-NN funcional, permitió identificar patrones temporales de deterioro. El enfoque temporalmente demostró su capacidad para anticipar insolvencias con alta sensibilidad, particularmente en el sector turismo colombiano.

Asimismo, se comprobó que es viable construir modelos de riesgo altamente competitivos sin sacrificar interpretabilidad. El clasificador desarrollado, aunque estructuralmente simple, alcanzó niveles de desempeño comparables a modelos avanzados como XGBoost o LSTM. Su ventaja crucial radica en la capacidad de explicar cada predicción tanto a través del peso específico de los indicadores financieros como mediante la analogía con empresas similares históricamente observadas. Esta doble capa de trazabilidad permite que el modelo sea comprensible por usuarios técnicos y no técnicos, alineándose con principios actuales de inteligencia artificial explicable y fortaleciendo su potencial de adopción práctica.

La investigación trasciende el plano teórico y propone una solución aplicable y replicable. El desarrollo de una aplicación interactiva basada en `Streamlit` (disponible en: https://tesis-knn.streamlit.app/), así como la publicación del código fuente en un repositorio abierto (https://github.com/tesisluisruizparedes/Tesis_KNN_Funcional), facilitan la transferencia del modelo a contextos reales de supervisión, auditoría o análisis sectorial. Esta viabilidad práctica, respaldada por una base de datos limpia y estructurada para empresas turísticas colombianas, refuerza el carácter utilizable de la

propuesta en escenarios donde se requiera monitoreo financiero continuo.

Bibliografía

Abrahamsen, N.-G. B., Nylén-Forthun, E., Møller, M., de Lange, P. E., & Risstad, M. (2024). Financial Distress Prediction in the Nordics: Early Warnings from Machine Learning Models [Open Access under CC BY 4.0]. *Journal of Risk and Financial Management*, *17*(10), 432. https://doi.org/10.3390/jrfm17100432

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 for tool selection. *Expert Systems with Applications*, *94*, 164-184. https://doi.org/10.1016/j.eswa.2017.10.040

Alam, N., Gao, J., & Jones, S. (2021). Corporate failure prediction: An evaluation of deep learning vs discrete hazard models. *Journal of International Financial Markets, Institutions & Money*, *75*, 101455. https://doi.org/10.1016/j.intfin.2021.101455

Alamari, M. B., Almulhim, F. A., Kaid, Z., & Laksaci, A. (2024). k-Nearest Neighbors Estimator for Functional Asymmetry Shortfall Regression. *Symmetry*, *16*(7), 928. https://doi.org/10.3390/sym16070928

Aljawazneh, H., Mora, A., García-Sánchez, P., & Castillo, P. (2021). Comparing the Performance of Deep Learning Methods to Predict Companies' Financial Failure. *IEEE Access*. https://doi.org/10.1109/ACCESS.2021.3093461

Almanjahie, I. M., Bouzebda, S., Kaid, Z., & Laksaci, A. (2024). The local linear functional kNN estimator of the conditional expectile: uniform consistency in number of neighbors. *Metrika*, *87*, 1-27. https://doi.org/10.1007/s00184-023-00942-0

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589-609.

Amer, A. A., Ravana, S. D., & Habeeb, R. A. A. (2025). Effective k-nearest neighbor models for data classification enhancement. *Journal of Big Data*, *12*(86). https://doi.org/10.1186/s40537-025-01137-2

Amirshahi, B., & Lahmiri, S. (2024). Bankruptcy prediction using optimal ensemble models under balanced and imbalanced data. *Expert Systems*, *41*(8), e13599. https://doi.org/10.1111/exsy.13599

Ansah-Narh, B., Yang, C.-H., & Lin, C.-W. (2024). Enhancing corporate bankruptcy prediction via a hybrid genetic algorithm and domain adaptation learning architecture. *Expert Systems with Applications*, *235*, 121289. https://doi.org/10.1016/j.eswa.2023.121289

Ansari, A., Ahmad, I. S., Abu Bakar, A., & Yaakub, M. R. (2020). A Hybrid Metaheuristic Method in Training Artificial Neural Network for Bankruptcy Prediction. *IEEE Access*, *8*, 176640-176650. https://doi.org/10.1109/ACCESS.2020.3026529

Antulov-Fantulin, N., Lagravinese, R., & Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior and Organization*, *183*, 681-699. https://doi.org/10.1016/j.jebo.2021.01.014

Antunes, F., Ribeiro, B., & Pereira, F. (2017). Probabilistic modeling and visualization for bankruptcy prediction. *Applied Soft Computing*, *60*, 831-843. https://doi.org/10.1016/j.asoc.2017.06.043

Baíllo, A., Cuevas, A., & Cuesta-Albertos, J. A. (2011). Supervised Classification for a Family of Gaussian Functional Models. *Scandinavian Journal of Statistics*, *38*(3), 480-498. https://doi.org/10.1111/j.1467-9469.2011.00734.x

Barboza, F., & Altman, E. (2024). Predicting financial distress in Latin American companies: A comparative analysis of logistic regression and random forest models. *North American Journal of Economics and Finance*, *72*, 102158. https://doi.org/10.1016/j.najef.2024.102158

Barboza, F., Cruz Basso, L. F., & Kimura, H. (2021). New metrics and approaches for predicting bankruptcy. *Communications in Statistics - Simulation and Computation*, 2615-2632. https://doi.org/10.1080/03610918.2021.1910837

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405-417. https://doi.org/10.1016/j.eswa.2017.04.006

Bargagli-Stoffi, F. J., Incerti, F., Riccaboni, M., & Rungi, A. (2023). Machine Learning for Zombie Hunting: Predicting Distress from Firms' Accounts and Missing Values [arXiv preprint]. https://arxiv.org/abs/2306.08165

Ben Jabeur, S., Gharib, C., Mefteh-Wali, S., & Ben Arfi, W. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*, 120658. https://doi.org/10.1016/j.techfore.2021.120658

Borges, A., & Carvalho, M. (2025). Short- and long-term financial distress prediction in SMEs: a survival model comparison [Special Issue: Computational Data Analysis and Numerical Methods (WCDANM 2024), Published online: 16 May 2025]. *Journal of Applied Statistics*. https://doi.org/10.1080/02664763.2025.2501166

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63*(6), 2899-2939.

Carmona, M., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: an extreme gradient boosting approach. *International Review of Economics & Finance*, *61*, 304-320. https://doi.org/10.1016/j.iref.2018.03.008

Carmona, P., Dwekat, A., & Mardawi, Z. (2022). No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. *Research in International Business and Finance*, *61*, 101649. https://doi.org/10.1016/j.ribaf.2022.101649

Chen, M. G. (2023). Estimating Sparsely and Irregularly Observed Multivariate Functional Data. *The ITEA Journal of Test and Evaluation*, *44*(3). https://doi.org/10.61278/itea.44.3.1009

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Chen, W., et al. (2025). Multi-class financial distress prediction based on hybrid feature selection and improved stacking ensemble model. *Expert Systems with Applications*.

Chen, W.-C., Chen, L.-C., Wang, C.-H., & Lee, C.-F. (2020). Ensemble learning with label proportions for bankruptcy prediction. *Information Sciences*, *516*, 108-121.

Chen, X., Liu, J., & Wu, C. (2025). Multi-class Financial Distress Prediction Based on Hybrid Feature Selection and Improved Stacking Ensemble

Model. *Expert Systems with Applications*, *282*, 127832. https://doi.org/10.1016/j.eswa.2025.127832

Cheraghali, M., & Molnár, J. (2024). SME default prediction: A systematic methodology-focused review. *Journal of Business Research*, *170*, 114199. https://doi.org/10.1080/00472778.2023.2277426

Chi, D.-J., & Shen, Z.-D. (2022). Using Hybrid Artificial Intelligence and Machine Learning Technologies for Sustainability in Going-Concern Prediction. *Sustainability*, *14*(3), 1810. https://doi.org/10.3390/su14031810

Climent, F., Momparler, A., & Carmona, M. (2019). Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach. *Journal of Financial Stability*, *40*, 100693. https://doi.org/10.1016/j.jbusres.2018.11.015

Correa, A. (2023). Predicting Business Bankruptcy in Colombian SMEs: A Machine Learning Approach. *Journal of International Commerce, Economics and Policy*, *14*(4), 2350027. https://doi.org/10.1142/S1793993323500278

Dasilas, A., & Rigani, A. (2024). Machine learning techniques in bankruptcy prediction: A systematic literature review. *Expert Systems with Applications*, *255*, 124761. https://doi.org/10.1016/j.eswa.2024.124761

du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research*, *254*, 236-252. https://doi.org/10.1016/j.ejor.2016.03.008

Erdogan, B. (2012). Prediction of bankruptcy using support vector machines. *International Journal of Computational Economics and Econometrics*, *2*(3), 241-252.

Feng, M., Shaonan, T., Lee, C., & Ma, L. (2019). Deep Learning Models for Bankruptcy Prediction Using Textual Disclosures. *European Journal of Operational Research*, *274*(2), 743-758. https://doi.org/10.1016/j.ejor.2018.10.024

Franco, Ó. J., Pérez, L., & Moreno, C. (2022). Análisis bibliométrico de la literatura sobre quiebra empresarial con enfoque en inteligencia artificial y aprendizaje automático. *Revista Colombiana de Ciencias Administrativas*, *40*(2), 21-45.

Galeano, P., et al. (2015). Mahalanobis distance for functional data with applications. *Wiley Interdisciplinary Reviews: Computational Statistics*.

Gnip, P., Kanász, R., Zoričak, M., & Drotár, P. (2025). An Experimental Survey of Imbalanced Learning Algorithms for Bankruptcy Prediction.

Artificial Intelligence Review, *58*, 104. https://doi.org/10.1007/s10462-025-11107-y

Guenani, S., Bouabsa, W., Fetitah, O., Kadi Attouch, M., & Khardani, S. (2024). Some Asymptotic Results of a kNN Conditional Mode Estimator for Functional Stationary Ergodic Data. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610926.2024.2384557

Guenani, S., Bouabsa, W., Kadi Attouch, M., & Fetitah, O. (2023). kNN Robustification Equivariant Nonparametric Regression Estimators for Functional Ergodic Data. *Hacettepe Journal of Mathematics and Statistics*, *52*(2), 512-528. https://doi.org/10.15672/hujms.1100871

Guerra, P., & Castelli, M. (2021). Machine Learning Applied to Banking Supervision: A Literature Review. *Risks*, *9*(7), 136. https://doi.org/10.3390/risks9070136

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horváthová, J., & Mokrišová, M. (2018). Risk of Bankruptcy, Its Determinants and Models [Open access under CC BY 4.0 license]. *Risks*, *6*(4), 117. https://doi.org/10.3390/risks6040117

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications*, *117*, 287-299. https://doi.org/10.1016/j.eswa.2018.09.039

Hu, W. (2022). k-NN nonparametric regression for negatively associated data with applications to bankruptcy prediction. *Journal of Computational and Applied Mathematics*, *400*, 113747. https://doi.org/10.1016/j.cam.2021.113747

Hu, X., Wang, J., Wang, L., & Yu, K. (2022). K-Nearest Neighbor Estimation of Functional Nonparametric Regression Model under NA Samples. *Axioms*, *11*(3), 102. https://doi.org/10.3390/axioms11030102

Huang, Y.-P., & Yen, M.-F. (2019). A New Perspective of Performance Comparison Among Machine Learning Algorithms for Financial Distress Prediction. *Applied Soft Computing*, *83*, 105663. https://doi.org/10.1016/j.asoc.2019.105663

Iparraguirre, J. A., Reinoso, D., & Pérez, M. (2024). Predicting business bankruptcy: A comparative analysis with machine learning models. *Expert Systems with Applications*, *235*, 121419. https://doi.org/10.1016/j.eswa.2023.121419

Issa, S., Bizel, G., Jagannathan, S. K., & Gollapalli, S. S. C. (2024). A Comprehensive Approach to Bankruptcy Risk Evaluation in the Financial Industry [Open access under CC BY 4.0 license]. *Journal of Risk and Financial Management*, *17*(1), 41. https://doi.org/10.3390/jrfm17010041

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* (2nd). Springer.

James, N., Menzies, M., & Chan, J. (2023). Semi-Metric Portfolio Optimization: A New Algorithm Reducing Simultaneous Asset Shocks. *Econometrics*, *11*(1), 8. https://doi.org/10.3390/econometrics11010008

Ji, H., Zhao, R., & Zhang, P. (2025). A dynamic financial risk prediction system for enterprises based on gradient boosting decision tree algorithm. *Expert Systems with Applications*, *231*, 120609. https://doi.org/10.1016/j.eswa.2023.120609

Kar, A. K., Akhter, M. M., Mishra, A. C., & Mohanty, S. K. (2024). EDMD: An Entropy based Dissimilarity measure to cluster Mixed-categorical Data. *Pattern Recognition*, *155*, 110674. https://doi.org/10.1016/j.patcog.2024.110674

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 3146-3154. https://papers.nips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Kim, S.-H., Lee, J.-Y., & Kim, Y.-K. (2022). Bankruptcy prediction using temporal data with long short-term memory models. *Mathematics*, *10*(9), 1345.

Kuizinienė, D., Krilavičius, T., Damaševičius, R., & Maskeliūnas, R. (2022). Systematic Review of Financial Distress Identification using Artificial Intelligence Methods. *Applied Artificial Intelligence*, *36*(1), 2138124. https://doi.org/10.1080/08839514.2022.2138124

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). *Applied Linear Regression Models* (4th). McGraw-Hill/Irwin.

Lin, F., Yeh, C.-C., & Lee, M.-Y. (2011). The Use of Hybrid Manifold Learning and Support Vector Machines in the Prediction of Business Failure. *Knowledge-Based Systems*, *24*(1), 95-101. https://doi.org/10.1016/j.knosys.2010.07.009

Lin, Y.-M., & Chang, P.-Y. (2023). Bankruptcy prediction using convolutional neural networks and SHAP interpretability. *Knowledge-Based Systems*, *258*, 110033. https://doi.org/10.1016/j.knosys.2022.110033

Liu, P., Yuan, H., Ning, Y., et al. (2024). A modified and weighted Gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses. *BMC Medical Research Methodology*, *24*(1), 305. https://doi.org/10.1186/s12874-024-02427-8

Liu, Y., Chen, H., & Zhang, W. (2025). Financial distress prediction with annual reports-based deep textual feature extraction: A hybrid approach. *Information Processing & Management*, *62*(2), 103574. https://doi.org/10.1016/j.ipm.2024.103574

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions [To appear in NeurIPS 2017]. *arXiv preprint arXiv:1705.07874*. https://doi.org/10.48550/arXiv.1705.07874

M., A., R. M., A. S., T. N., D., S., M. S., & J., V. (2025). Predicting Bankruptcy With Precision: Insights From Hybrid Machine Learning Models On Unbalanced Polish Financial Data. *International Journal for Scientific Research and Technology (IJSART)*, *11*(5), 57-65. https://ijsart.com/public/storage/paper/pdf/IJSARTV11I5103465.pdf

Marso, K., Malebary, S. J., & Belkacem, S. (2020). Bankruptcy prediction using hybrid neural networks with artificial bee colony. *Computational Intelligence and Neuroscience*.

Muslim, M. A., & Yosza, D. (2021). Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning. *International Journal of Electrical and Computer Engineering*, *11*(6), 5549-5557. https://doi.org/10.11591/ijece.v11i6.pp5549-5557

Nagendrakumar, N., Alwis, K. N. N., Eshani, U. A. K., & Kaushalya, S. B. U. (2023). Prediction of Corporate Financial Distress in the Travel and Tourism Industry [Open access under CC BY 4.0 license]. *Corporate Ownership & Control*, *20*(3 (Special Issue)), 262-267. https://doi.org/10.22495/cocv20i3siart2

Nayak, S. M., & Rout, M. (2023). A predictive model for bankruptcy: ANN, LSTM and CNN approaches. *Journal of Statistics and Management Systems*, *26*(1), 67-86. https://doi.org/10.47974/JSMS-948

Nazareth, D. L., & Reddy, R. (2023). Applications of Machine and Deep Learning in Finance: A Systematic Review. *Applied Artificial Intelligence*, *37*(1), 2162678. https://doi.org/10.1016/j.eswa.2023.119640

Ó Searcóid, M. (2007). *Metric Spaces*. Springer.

Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *Proceedings of the International Joint Conference on Neural Networks*, 163-168.

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

Papík, M., & Papíková, L. (2024a). Automated Machine Learning in Bankruptcy Prediction of Manufacturing Companies. *Procedia Computer Science*, *232*, 1428-1436. https://doi.org/10.1016/j.procs.2024.01.141

Papík, M., & Papíková, L. (2024b). Automated Machine Learning in Bankruptcy Prediction of Manufacturing Companies. *Procedia Computer Science*, *232*, 1428-1436. https://doi.org/10.1016/j.procs.2024.01.141

Papík, M., Papíková, L., Kajanová, J., & Bečka, M. (2022). CatBoost: The case of bankruptcy prediction. En *Sustainable Finance, Digitalization and the Role of Technology* (pp. 3-17, Vol. 487). Springer. https://doi.org/10.1007/978-3-031-08084-5_3

Park, S., Kim, D.-S., & Kim, S. J. (2021). Explainability of Machine Learning Models for Bankruptcy Prediction. *Journal of Risk and Financial Management*, *14*(12), 535.

Pellegrino, A., Ruiz, D., & Conti, M. (2024). Multi-head LSTM architecture for corporate bankruptcy prediction [In press]. *Journal of Forecasting*.

Qian, H., Wang, B., Yuan, M., Gao, S., & Song, Y. (2022). Financial Distress Prediction Using a Corrected Feature Selection Measure and Gradient Boosted Decision Tree. *Expert Systems with Applications*, *190*, 116202. https://doi.org/10.1016/j.eswa.2021.116202

Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *162*, 895-899. https://doi.org/10.1016/j.procs.2019.12.065

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv preprint arXiv:1602.04938*. https://doi.org/10.48550/arXiv.1602.04938

Romero Espinosa, F., Melgarejo Molina, Z. A., & Vera-Colina, M. A. (2015). Fracaso empresarial de las pequeñas y medianas empresas (pymes) en Colombia. *SUMA de Negocios*, *6*(13), 29-41. https://doi.org/10.1016/j.sumneg.2015.08.003

Rudin, W. (1976). *Principles of Mathematical Analysis* (3rd). McGraw-Hill.

Santiago, K., Yanes, A., & Mercado-Caruso, N. (2024). Analyzing Correlations in Sustainable Tourism Perception: Statistical Insights from Diverse

Caribbean Colombian Tourist Sites. *Procedia Computer Science*, *231*, 490-495. https://doi.org/10.1016/j.procs.2023.12.239

Shetty, S., Musa, M., & Brédart, X. (2022). Bankruptcy Prediction Using Machine Learning Techniques. *Journal of Risk and Financial Management*, *15*(1), 35. https://doi.org/10.3390/jrfm15010035

Shi, Y., & Li, X. (2019). A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms. *Heliyon*, *5*(7), e02997. https://doi.org/10.1016/j.heliyon.2019.e02997

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, *74*(1), 101-124.

Son, H., Hyun, C., Phan, D., & Hwang, H. (2019). Data Analytic Approach for Bankruptcy Prediction. *Expert Systems with Applications*, *138*, 112816. https://doi.org/10.1016/j.eswa.2019.07.033

Song, Y., Chiangpradit, M., & Busababodhin, P. (2025). Hyperband-Optimized CNN-BiLSTM with Attention Mechanism for Corporate Financial Distress Prediction. *Applied Sciences*, *15*(11), 5934. https://doi.org/10.3390/app15115934

Superintendencia de Sociedades. (2023, marzo). Informe Económico-Financiero del Sector Turismo 2019–2021 [Elaborado por el Grupo de Estudios Empresariales. Bogotá, Colombia.].

Téllez, J., Hernández, M., & Murcia, M. P. (2024, octubre). Colombia: Situación Turismo [Fecha de cierre: 18 de octubre de 2024]. https://www.bbvaresearch.com/publicaciones/colombia-situacion-turismo-noviembre-2024/

Thompson, J. R., & Davison, M. (2024). Functional Mixed-type Clustering of Investors' Daily Returns During a Market Shock Change-point and Recovery. *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24)*. https://doi.org/10.1145/3677052.3698633

Tian, X., & Yu, J. (2012). Recent advances on support vector machines research. *Neurocomputing*, *74*(1-3), 1-6.

Tsai, C.-F., Hsu, Y.-F., Yen, D. C., & Hsu, C.-C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, *24*, 977-984.

Tserng, H.-C., Lin, C.-Y., & Chen, P.-H. (2011). An enforced support vector machine model for construction contractor default prediction. *Automation in Construction*, *20*(8), 1242-1249.

Vallarino, D. (2024). A Comparative Machine Learning Survival Models Analysis for Predicting Time to Bank Failure in the US (2001–2023)

[Open access under CC BY 4.0 license]. *Journal of Economic Analysis*, *3*(1), 129-144. https://doi.org/10.58567/jea03010007

Wang, G., Ma, J., Xu, Y., & Wang, H. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Computers & Industrial Engineering*, *74*, 50-61.

Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application*, *3*, 257-295. https://doi.org/10.1146/annurev-statistics-041715-033624

Wang, S., & Chi, G. (2024). Cost-sensitive stacking ensemble learning for company financial distress prediction. *Expert Systems with Applications*, *255*(Part A), 124525. https://doi.org/10.1016/j.eswa.2024.124525

Wang, X., & Brorsson, M. (2024). Augmenting Bankruptcy Prediction Using Reported Behavior of Corporate Restructuring. *International Conference on Information and Communication Technology (IC 2023)*, *2036*, 1-20. https://doi.org/10.1007/978-981-97-0065-3_8

Wilson, R. L., & Sharda, R. (1994). A comparison of neural network and logistic regression predictions. *Decision Support Systems*, *11*(5), 545-557.

Yousaf, U. B., Jebran, K., & Wang, M. (2022). A Comparison of Static, Dynamic and Machine Learning Models in Predicting the Financial Distress of Chinese Firms. *Romanian Journal of Economic Forecasting*, *25*(1), 122-138.

Zapata, A., & Mukhopadhyay, S. (2024). Evaluation of Hybrid Models for Bankruptcy Prediction [Citado en *2024_SME default prediction A systematic methodology-focused review*]. En *SME Default Prediction: A Systematic Methodology-Focused Review*. Springer.

Zelenkov, Y., Fedorova, E., & Chekrizov, D. (2017). Two-step classification method based on genetic algorithm for bankruptcy forecasting. *Expert Systems with Applications*, *88*, 393-401. https://doi.org/10.1016/j.eswa.2017.07.025

Zhang, H., Zhang, Q., & Li, Y. (2016). A nonlinear subspace multiple kernel learning for financial distress prediction of Chinese listed companies. *Knowledge-Based Systems*, *94*, 102-111.

Zhang, H., & Zhang, W. (2025). Advancing enterprise risk management with deep learning: A predictive approach using the XGBoost-CNN-BiLSTM model. *PLOS ONE*, *20*(4), e0319773. https://doi.org/10.1371/journal.pone.0319773

Zhao, J., Ouenniche, J., & De Smedt, J. (2024a). A complex network analysis approach to bankruptcy prediction using company relational information-

based drivers. *Knowledge-Based Systems*, *300*, 112234. https://doi.org/10.1016/j.knosys.2024.112234

Zhao, J., Ouenniche, J., & De Smedt, J. (2024b). Survey, classification and critical analysis of the literature on corporate bankruptcy and financial distress prediction. *Machine Learning with Applications*, *15*, 100527. https://doi.org/10.1016/j.mlwa.2024.100527

Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, *58*, 93-101. https://doi.org/10.1016/j.eswa.2016.04.001