**Maestría en**

# CIENCIA DE DATOS Y MAQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

**Trabajo previo a la obtención de título de Magister en Ciencia de Datos y Máquinas de Aprendizaje con Mención en Inteligencia Artificial**


**AUTORES:**

Roberth Gabriel Lima Carvajal

Rubén Darío Mena Nazca

María Estefanía Sánchez Pacheco

Fernando José Zambrano Farías

**TUTORES:**

Karla Estefanía Mora Cajas

Héctor Guillermo Avalos Silva


Predicción del desempeño financiero futuro de empresas PYMES ecuatorianas mediante técnicas de aprendizaje automático.


Quito - Ecuador

Julio – 2026


REINVENTEMOS EL FUTURO

## Certificación de autoría

Nosotros, **Roberth Gabriel Lima Carvajal, Rubén Darío Mena Nazca, María Estefanía Sánchez Pacheco, Fernando José Zambrano Farías,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
ROBERTH GABRIEL LIMA CARVAJAL

---

**Firma**
**Roberth Gabriel Lima Carvajal**

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
RUBEN DARIO MENA NAZCA

---

**Firma**
**Rubén Darío Mena Nazca**

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
MARIA ESTEFANIA SANCHEZ PACHECO

---

**Firma**
**María Estefanía Sánchez Pacheco**

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
FERNANDO JOSE ZAMBRANO FARIAS

---

**Firma**
**Fernando José Zambrano Farías**

## Autorización de Derechos de Propiedad Intelectual

Nosotros, **Roberth Gabriel Lima Carvajal, Rubén Darío Mena Nazca, María Estefanía Sánchez Pacheco y Fernando José Zambrano Farías**, en calidad de autores del trabajo de investigación titulado ***Predicción del desempeño financiero futuro de empresas PYMES ecuatorianas mediante técnicas de aprendizaje automático***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, julio de 2026

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
ROBERTH GABRIEL
LIMA CARVAJAL

---

**Firma**
**Roberth Gabriel Lima Carvajal**

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
RUBEN DARIO MENA
NAZCA

---

**Firma**
**Rubén Darío Mena Nazca**

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
MARIA ESTEFANIA
SANCHEZ PACHECO

---

**Firma**
**María Estefanía Sánchez Pacheco**

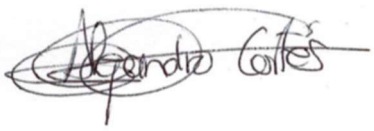
Validar únicamente en FirmaEC.
Firmado electrónicamente por:
FERNANDO JOSE
ZAMBRANO FARIAS

---

**Firma**
**Fernando José Zambrano Farías**

**Aprobación de dirección y coordinación del programa**

Nosotros, **Director Alejandro Cortés López de EIG y Coordinadora Karla Estefanía Mora Cajas de UIDE,** declaramos que: **Roberth Gabriel Lima Carvajal, Rubén Darío Mena Nazca, María Estefanía Sánchez Pacheco y Fernando José Zambrano Farías** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



---

**Alejandro Cortés López**
Director/a de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

---

**Karla Estefanía Mora Cajas**
Coordinador/a de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

## AGRADECIMIENTOS

En primer lugar, expresamos nuestro más profundo agradecimiento a la Universidad Internacional del Ecuador (UIDE) y a la Escuela de Inteligencia y Gestión (EIG) por abrirnos las puertas a este programa de vanguardia y brindarnos las herramientas necesarias para desarrollarnos en el fascinante campo de la Ciencia de Datos.

A nuestros tutores, cuyo conocimiento, guía constante y paciencia fueron pilares fundamentales para direccionar metodológicamente esta investigación y transformar las ideas en un pipeline analítico robusto y reproducible. De igual manera, agradecemos a la Superintendencia de Compañías, Valores y Seguros (SCVS) por facilitar el acceso a la información oficial que sirvió como base empírica para este estudio.

A nuestros compañeros de equipo: Rubén, María Estefanía, Fernando y mi persona; gracias por las extenuantes jornadas de programación, discusiones metodológicas y el apoyo mutuo indispensable para culminar con éxito este desafío académico.

Finalmente, un agradecimiento muy especial a nuestras familias, amigos y seres queridos, quienes con su comprensión, paciencia y aliento incondicional sostuvieron nuestro esfuerzo durante este trayecto de especialización profesional.

*Roberth Lima*

A Díos por ser mi guía en cada etapa de mi vida.

Agradezco profundamente a la Universidad Internacional del Ecuador y a sus líderes por haber confiado en mí, permitiéndome crecer y formarme como profesional tanto en lo académico como en lo laboral, y por ser un espacio que ha marcado mi desarrollo personal y profesional.

A mis compañeros: María Estefanía, Fernando, Roberth; cuyo aporte y experiencias, fueron fundamentales para ampliar mi visión y hacer posible el desarrollo de este trabajo de titulación.

*Darío Mena*

Mi profundo agradecimiento a mis compañeros: Roberth, Darío y Estefanía por el aporte de cada uno en este trabajo.

*Fernando Zambrano*

Agradezco a Dios por permitirme vivir esta nueva experiencia. Gracias compañeros: Roberth, Ruben y Fernando por el trabajo en conjunto.

Estefanía Sánchez

## RESUMEN

Las pequeñas empresas representan un motor fundamental para la economía ecuatoriana, concentrando más del 90 % del tejido empresarial formal. Sin embargo, su vulnerabilidad ante las crisis económicas hace necesario contar con herramientas que ayuden a anticipar su comportamiento financiero. Por esta razón, el presente trabajo desarrolla un modelo analítico basado en aprendizaje automático para predecir si una empresa será rentable en el futuro cercano.

Para lograrlo, utilizamos los estados financieros entregados a la Superintendencia de Compañías, Valores y Seguros (SCVS) entre 2020 y 2024. El problema se abordó como una clasificación binaria, donde la información contable de un año ($t$) nos sirve para pronosticar la situación del año siguiente ($t + 1$). Como es común que en este tipo de datos existan muchos más casos de empresas rentables que de empresas con problemas (desbalance de clases), aplicamos la técnica SMOTE+Tomek para equilibrar la información antes de entrenar a los algoritmos.

En el desarrollo del proyecto, comparamos tres modelos: Regresión Logística, Random Forest y XGBoost. También usamos K-Means y DBSCAN para agrupar a las empresas y entender mejor las diferencias que existen entre ellas. Los resultados demostraron que XGBoost funciona mejor con los datos originales, pero la Regresión Logística logra el mejor desempeño una vez que los datos se balancean con SMOTE+Tomek. Además, la etapa de segmentación nos permitió ver con claridad qué factores influyen más en el éxito financiero de las PYMES.

Al final, esta herramienta resulta de gran utilidad práctica para tomar decisiones estratégicas, mejorar la evaluación de créditos y orientar políticas de apoyo empresarial.

.


**Palabras clave:** aprendizaje automático, desempeño financiero, pequeñas empresas, modelos predictivos, segmentación empresarial, SMOTE+Tomek

## ABSTRACT

Small businesses are a fundamental driver for the Ecuadorian economy, accounting for more than 90% of the formal business fabric. However, their vulnerability to economic crises makes it necessary to have tools that help anticipate their financial behavior. For this reason, this study develops an analytical model based on machine learning to predict whether a company will be profitable in the near future.

To achieve this, we used the financial statements submitted to the Superintendence of Companies, Securities, and Insurance (SCVS) between 2020 and 2024. The problem was approached as a binary classification, where the accounting information from one year ($t$) helps us forecast the situation for the following year ($t + 1$). Since it is common with this type of data to have many more cases of profitable companies than troubled ones (class imbalance), we applied the SMOTE+Tomek technique to balance the information before training the algorithms.

During the project's development, we compared three models: Logistic Regression, Random Forest, and XGBoost. We also used K-Means and DBSCAN to group the companies and better understand the differences that exist among them. The results demonstrated that XGBoost works best with the original data, but Logistic Regression achieves the best performance once the data is balanced with SMOTE+Tomek. Additionally, the segmentation stage allowed us to clearly see which factors have the greatest influence on the financial success of SMEs.

Ultimately, this tool is highly useful in practice for making strategic decisions, improving credit evaluations, and guiding business support policies.

**Keywords:** machine learning, financial performance, small businesses, predictive models, business segmentation, SMOTE+Tomek.

# CONTENIDO

# Lista de tablas

# Lista de figuras

# CAPITULO 1

## 1. INTRODUCCION

### 1.1. Planteamiento del problema e importancia del estudio

Las pequeñas empresas constituyen uno de los pilares fundamentales de las economías emergentes debido a su contribución a la generación de empleo, la diversificación de la oferta productiva y la dinamización de los mercados locales (Dieperink et al., 2025; Sánchez Pacheco et al., 2025). En América Latina, las micro, pequeñas y medianas empresas (mipymes) representan aproximadamente el 99 % del tejido empresarial y constituyen un componente fundamental para la generación de empleo y el desarrollo económico de la región (CEPAL, 2009). En Ecuador, este segmento participa en diversas actividades productivas y contribuye al desarrollo económico de distintas regiones del país.

A pesar de su relevancia, este segmento enfrenta una elevada vulnerabilidad financiera, caracterizada por restricciones de acceso al financiamiento, limitaciones en inversiones tecnológicas y una menor capacidad para absorber choques económicos adversos (INEC, 2017b). Estas dificultades comprometen la sostenibilidad de las empresas y generan impactos en el empleo y la recaudación tributaria

En Ecuador, esta situación toma una especial relevancia. Las pequeñas empresas realizan sus actividades en un ambiente macroeconómico marcado por incertidumbre económica, fuerte variabilidad en la demanda, cambios constantes en los marcos regulatorios y diversidad en los sectores productivos. Estas circunstancias influyen directamente en su capacidad para sostener niveles apropiados de rentabilidad y garantizar su sostenibilidad en el mercado empresarial. La evidencia empírica en el país muestra que las dificultades financieras de este segmento no solo comprometen la continuidad de estas compañías, sino que también repercuten en la generación de empleo, la recaudación tributaria y el crecimiento en la economía ecuatoriana (INEC, 2017a; Zambrano-Farías et al., 2023).

Tradicionalmente, la vasta literatura sobre fracaso empresarial se ha concentrado en la identificación anticipada de escenarios de insolvencia financiera. Desde los primeros estudios de Altman (1968) y Ohlson (1980), numerosos trabajos han demostrado que ciertos indicadores financieros representan señales que son capaces de anticipar escenarios de vulnerabilidad. Sin embargo, gran parte de estas investigaciones se ha orientado hacia la predicción del fracaso empresarial una vez que los problemas financieros comienzan a manifestarse, prestando menor atención a la estimación del desempeño financiero futuro como un fenómeno continuo y preventivo.

Bajo un escenario de administración financiera, anticipar la capacidad de una empresa para generar resultados positivos en el período siguiente resulta tan importante como identificar circunstancias de insolvencia financiera. La rentabilidad corporativa futura es una señal del desempeño financiero (Sánchez Pacheco et al., 2022), al evidenciar al mismo tiempo la

eficiencia operativa, la estructura financiera y la capacidad de adaptación de la empresa frente a su entorno. Por ende, la posibilidad de predecir el desempeño financiero futuro a partir de información pasada de la empresa representa una herramienta para tomar decisiones estratégicas, hacer una evaluación crediticia y diseñar políticas que permitan el fortalecimiento empresarial (Valls Martínez et al., 2026).

A pesar de que la Superintendencia de Compañías, Valores y Seguros dispone de información contable y financiera detallada y de cobertura nacional, la construcción de modelos predictivos enfocados específicamente a pequeñas empresas ecuatorianas continúa siendo limitada (Zambrano-Farías et al., 2022). En la mayoría de investigaciones se han replicado modelos estadísticos construidos para economías con empresas diferentes o se han utilizado enfoques centrados en la explicación del desempeño pasado, sin explotar plenamente el potencial predictivo de las técnicas contemporáneas de ciencia de datos (Rafatnia et al., 2020; Ragab & Saleh, 2021; Vašaničová et al., 2025).

En los últimos años, el desarrollo de la ciencia de datos y técnicas de *machine learning* han cambiado la manera en que se analizan los fenómenos financieros. Algoritmos supervisados como la Regresión Logística, *Random Forest* y *XGBoost* permiten identificar patrones complejos contenidos en grandes volúmenes de información y estimar la probabilidad de ocurrencia de eventos futuros con niveles elevados de precisión (Chen & Guestrin, 2016; Hastie et al., 2017). Adicionalmente, gracias a las técnicas no supervisadas es posible identificar perfiles empresariales similares, revelando estructuras latentes que no son evidentes mediante enfoques tradicionales de clasificación.

Aunque el estudio de la analítica de datos ha avanzado, no todas las empresas los aprovechan de la misma manera. Las grandes compañías tienen los recursos y los equipos de expertos necesarios para crear modelos de predicción, pero las pequeñas empresas viven una realidad distinta: tienen poca información organizada, presupuestos limitados y falta de personal técnico. Por esta razón, la mayoría de las decisiones en estos negocios se siguen tomando con base en la experiencia de los administradores o usando los indicadores financieros básicos de siempre.

Al ver esta diferencia, se evidenció de que es necesario investigar cómo funcionan los algoritmos de aprendizaje automático cuando se aplican a empresas más pequeñas, específicamente en el Ecuador. La idea de analizar este escenario es comprobar si los modelos siguen siendo precisos al trabajar con bases de datos reales de nuestro país, ya que la información financiera de nuestras PYMES es muy diversa y varía mucho de una empresa a otra.

Además de predecir el desempeño, este estudio también busca aplicar técnicas de segmentación para ver si podemos agrupar a las empresas y encontrar perfiles financieros distintos entre ellas. Por último, queremos evaluar si el uso de estrategias para equilibrar los datos, como la técnica SMOTE+Tomek, realmente ayuda a que los modelos funcionen mejor en nuestro entorno.

A pesar de todo esto, en Ecuador todavía falta en descubrir qué tan útiles son realmente estas herramientas para predecir el futuro financiero de las pequeñas empresas. Esto es aún más importante si tomamos en cuenta que nuestro estudio abarca un periodo muy inusual, marcado por la pandemia de COVID-19 y los años de recuperación económica. Tampoco hay mucha

información previa sobre cómo afecta el desbalance de los datos a las predicciones de los modelos, ni sobre las verdaderas diferencias que existen internamente entre los negocios de este sector.

En este contexto, surge la siguiente pregunta de investigación:

**¿Es posible predecir el desempeño financiero futuro de las pequeñas empresas ecuatorianas mediante técnicas de aprendizaje automático, utilizando indicadores financieros del ejercicio fiscal anterior reportados a la Superintendencia de Compañías, Valores y Seguros durante el período 2020–2024?**

La respuesta a esta interrogante posee implicaciones tanto académicas como prácticas. Desde el punto de vista científico, contribuye al desarrollo de evidencia empírica sobre la aplicación de técnicas de ciencia de datos en economías emergentes. Desde una perspectiva aplicada, proporciona herramientas orientadas a fortalecer la capacidad de anticipación de empresarios, entidades financieras y organismos públicos, promoviendo decisiones preventivas sustentadas en evidencia cuantitativa.

## 1.2. Definición del proyecto

El objetivo de este trabajo de investigación que representa un trabajo de titulación es diseñar, desarrollar y validar un modelo de predicción del desempeño financiero futuro de las pequeñas empresas ecuatorianas mediante técnicas de *machine learning* (aprendizaje automático), para ello se información proveniente de estados financieros que fueron reportados a la Superintendencia de Compañías, Valores y Seguros (SCVS) durante el período 2020–2024. El estudio se desarrolló con datos de empresas a nivel nacional y que pertenecen a diversos sectores económicos, de acuerdo con la Clasificación Nacional de Actividades Económicas (CIIU Rev. 4.0).

Al contrario de lo que señala la literatura tradicional, que suele limitarse a identificar el fracaso empresarial o la falta de liquidez cuando estos problemas son visibles, el estudio tomó una posición estrictamente preventiva. El interés es proyectar el desempeño financiero de las compañías en el periodo futuro inmediato. Desde una perspectiva metodológico, se configuró este análisis como un modelo de clasificación supervisada. Así, la variable que se va a predecir es el estado económico que alcanzará la empresa en el siguiente ejercicio fiscal ($t + 1$), tomando como punto de partida la información contable disponible en el año en curso ($t$).

Desde una perspectiva teórica, el estudio se sitúa entre la literatura sobre predicción financiera y los desarrollos contemporáneos en ciencia de datos aplicada a las finanzas, es decir, considera los fundamentos clásicos de la predicción del fracaso empresarial que fueron planteados por Altman (1968) y Ohlson (1980), quienes demostraron que la información contable tiene la capacidad de anticipar el comportamiento futuro de las organizaciones. Ahora bien, integra estos aportes dentro de un enfoque actualizado que incorpora herramientas de *machine learning* encaminadas no solamente a clasificar observaciones, sino también a identificar tendencias patrones complejos presentes cuando se trabaja con grandes volúmenes de datos.

Para lograr este análisis, el diseño metodológico compara tres algoritmos supervisados de distinta naturaleza: Regresión Logística, *Random Forest y XGBoost*. El objetivo de comparar estos tres modelos es comprobar empíricamente si las herramientas de mayor complejidad computacional superan la capacidad predictiva de los métodos probabilísticos clásicos. Adicionalmente, dado el sesgo que tiene la información financiera, esta investigación evalúa el impacto de la técnica híbrida SMOTE+Tomek. La intención es verificar si el remuestreo de los datos realmente ayuda a los modelos a detectar con mayor eficacia la clase minoritaria.

En su segunda fase, el estudio no se restringe solamente a modelos de predicción, sino que además integra enfoques de aprendizaje no supervisado, utilizando específicamente K-Means y DBSCAN como técnicas de clasificación. El objetivo de esta etapa consiste en identificar la amplia diversidad interna del sector empresarial evaluado. Gracias a este enfoque, es posible revelar estructuras latentes y perfiles financieros muy definidos que pasarían completamente desapercibidos bajo un esquema exclusivo de clasificación binaria.

Con respecto a las variables consideradas predictoras o explicativas, en el estudio se consideró la construcción de un conjunto de indicadores derivados del estado de situación financiera y del estado de resultados, incluyendo factores relacionados con la rentabilidad, la eficiencia operativa, la estructura de los gastos de la empresa, el nivel de solvencia de largo plazo o endeudamiento, el tamaño de la compañía y la antigüedad de la empresa en el sector en el que desarrolla sus actividades. La selección de estas variables se sustentó tanto en la evidencia empírica previa como en su capacidad potencial para explicar el desempeño financiero futuro (Durica et al., 2019; Parra, 2011; Romero Espinosa, 2013; Zambrano-Farías et al., 2024).

Finalmente, el estudio incorporó procedimientos de interpretabilidad para identificar los factores que más destacan al momento de explicar el desempeño financiero futuro, de esta manera se transforman los resultados en evidencia útil para la toma de decisiones empresariales y la formulación de estrategias preventivas encaminadas a fortalecer la sostenibilidad financiera de las pequeñas empresas ecuatorianas en el largo plazo.

### 1.3. Naturaleza o tipo de proyecto

La presente investigación adoptó un enfoque cuantitativo ya que el fenómeno estudiado puede representarse mediante variables numéricas provenientes de los estados financieros reportados por las empresas objeto de estudio (Hernández Sampieri et al., 2014). El análisis se sustentó en información objetiva y verificable, lo que permitió identificar patrones observables, estimar relaciones entre variables y desarrollar modelos predictivos basados en evidencia empírica.

Este trabajo de investigación tiene un alcance predictivo, orientado a estimar la probabilidad de ocurrencia de futuros escenarios mediante técnicas de *machine learning* (aprendizaje automático). Asimismo, incluye elementos descriptivos y analíticos: por un lado, expone la evolución de los principales indicadores financieros de las pequeñas empresas

durante el período 2020–2024; por otro, evalúa cómo diferentes variables contribuyen a explicar el desempeño financiero futuro y compara el comportamiento de distintos algoritmos bajo escenarios de desbalance de clases.

En lo que respecta al diseño metodológico, la investigación fue no experimental, puesto que no se manipularon las variables o atributos de forma deliberada (Hernández Sampieri et al., 2014). La información que se analizó provino de registros financieros generados en el ejercicio fiscal de la actividad de la compañía y que fueron reportados oficialmente a la SCVS. Asimismo, los datos de son de corte longitudinal al incorporar observaciones correspondientes a ejercicios fiscales consecutivos comprendidos entre 2020 y 2024.

Desde la aplicación, este trabajo de fin de maestría adopta los fundamentos de la ciencia de datos. El diseño metodológico consolida etapas de preprocesamiento, modelado supervisado, evaluación predictiva, interpretabilidad y segmentación dentro de un *pipeline* completamente reproducible, estructurado sobre *notebooks* computacionales. Esta arquitectura asegura un alto rigor académico sin sacrificar su viabilidad en el mundo real, transformando los algoritmos en recursos de apoyo directo para el sector empresarial, las instituciones de crédito y los entes reguladores.

Finalmente, la concentración de herramientas de aprendizaje supervisado y no supervisado mejora notablemente la lectura del mercado. Por un lado, las técnicas de clasificación utilizadas actúan como un motor para pronosticar la rentabilidad a futuro. Por el otro, las estrategias de agrupamiento muestran perfiles de empresas muy singulares. Toda esta información demuestra que las pequeñas empresas en Ecuador no conforman un grupo con características semejantes, sino que operan bajo dinámicas financieras totalmente diferentes.

## 1.4. Objetivos

### *1.4.1.* Objetivo general

Desarrollar un modelo de predicción, basado en técnicas de *machine learning* (aprendizaje automático), que permita medir el desempeño financiero futuro de pequeñas empresas en Ecuador a partir de su información contable y financiera proveniente del ejercicio fiscal anterior, utilizando información reportada por la Superintendencia de Compañías, Valores y Seguros (SCVS) durante el período 2020–2024.

### *1.4.2.* Objetivos específicos

- Construir una base de datos financiera de pequeñas empresas ecuatorianas a partir de fuentes oficiales, aplicando criterios de consistencia contable y segmentación por tamaño empresarial.
- Identificar perfiles financieros homogéneos mediante técnicas de agrupamiento no supervisado, específicamente *K-Means* y *DBSCAN*.

- Comparar modelos de clasificación supervisada, incluyendo regresión logística, *Random Forest* y *XGBoost*, para predecir el desempeño financiero del ejercicio $t+1$ a partir de variables financieras observadas en el ejercicio $t$.
- Evaluar el impacto del desbalance de clases sobre el desempeño predictivo mediante la comparación entre modelos entrenados con datos originales y modelos balanceados utilizando la técnica híbrida *SMOTE+Tomek*.
- Determinar los grupos empresariales con mayor vulnerabilidad financiera integrando los resultados de la segmentación y los factores identificados por los modelos predictivos.

## 1.5. Justificación e importancia del trabajo de investigación

La sostenibilidad operativa y financiera de las pequeñas empresas representan un factor decisivo para la estabilidad económica y social de las economías emergentes (Valls Martínez et al., 2026). En Ecuador, este conglomerado de empresas representa un porcentaje importante del tejido empresarial formal y desempeña un papel fundamental en la generación de empleo, la articulación de cadenas productivas y la dinamización de los mercados locales. A pesar de ello, estas empresas se enfrentan a muchas restricciones de acceso al financiamiento, limitaciones tecnológicas y una menor capacidad para resistir crisis e inestabilidad económica, lo que incrementa su vulnerabilidad a períodos de bajo desempeño financiero (Adekunle, 2011; INEC, 2017b).

Debido la importancia de este grupo de empresas dentro de una economía, la necesidad de proyectar el comportamiento económico de estas compañías va mucho más allá del simple interés corporativo. Cuando la rentabilidad de las pequeñas empresas disminuye, el impacto afecta directamente a la estabilidad laboral, merma los ingresos tributarios y frena el desarrollo de comunidades enteras. Frente a este escenario económico, el diseño de modelos estadísticos de predicción tiene una notable importancia, ya que contar con señales anticipadas brinda un margen de maniobra necesario para proteger la sostenibilidad de estos negocios ante escenarios adversos.

Bajo esta perspectiva, el pronosticar las fluctuaciones contables representa una ventaja estratégica invaluable. Identificar los primeros síntomas de insolvencia financiera permite a los administradores la oportunidad de aplicar medidas correctivas y estratégicas mucho antes de que la crisis alcance un punto de no retorno. Más allá de las paredes de la organización, estas proyecciones inyectan certidumbre al entorno de la empresa; bancos, inversionistas e instituciones gubernamentales ganan un respaldo técnico superior para medir el riesgo y encauzar los recursos de forma eficiente. El esfuerzo por mejorar la exactitud de estos modelos abandona el terreno de la teoría para convertirse en una solución práctica, impulsando un entorno de negocios donde las decisiones se toman con base en evidencia empírica.

Con frecuencia la investigación sobre la predicción del desempeño empresarial, se ha centrado en la identificación de factores que inciden en el fracaso empresarial o la insolvencia financiera cuando los síntomas de deterioro ya han ocurrido. Los modelos clásicos de Altman (1968) y Ohlson (1980) concluyeron que la información contable y financiera es de mucha

utilidad para anticipar dificultades futuras. Posteriormente, la incorporación de modelos probabilísticos y, más recientemente, de algoritmos de *machine learning*, incrementó las posibilidades analíticas disponibles para analizar problemas complejos de clasificación financiera (Durica et al., 2019; Pietrzak, 2022).

Pese a la abundancia de literatura sobre predicción corporativa, la inmensa mayoría de estos estudios basa sus hallazgos en países que tienen mercados desarrollados. En Ecuador, por ejemplo, la escasez de investigaciones enfocadas en la pequeña empresa obliga a replicar herramientas predictivas que ignoran la dinámica local. Esta dependencia teórica crea un vacío crítico: el país cuenta con abundantes registros financieros oficiales, pero carece de modelos predictivos calibrados para entender su propia realidad.

Utilizar herramientas analíticas usadas en otros contextos de manera directa conlleva una posibilidad de sesgo elevado. Las variables que rigen el ecosistema productivo nacional (como las severas barreras crediticias, los choques macroeconómicos o los hábitos contables de las pymes) alteran radicalmente el peso de los indicadores. Frente a este panorama, validar los algoritmos de clasificación exclusivamente con datos autóctonos deja de ser una opción y se convierte en un requisito indispensable. Solo así, los resultados de la modelación reflejarán las reglas del mercado interno en lugar de confirmar teorías de contextos foráneos.

Para acortar esta diferencia, esta investigación procesa la base de datos de la Superintendencia de Compañías, Valores y Seguros a lo largo del periodo 2020-2024. El valor de este horizonte de tiempo radica en su altísima volatilidad, pues captura tanto el colapso operativo provocado por la pandemia de COVID-19 como los esfuerzos posteriores de reactivación. El analizar y evaluar esta información somete a los algoritmos a un entorno de estrés e incertidumbre, un escenario ideal para poner a prueba el verdadero aguante del tejido empresarial.

Con lo que respecta a la metodología, el diseño del estudio integra diferentes técnicas de ciencia de datos a través de un flujo de procesamiento reproducible. Al contrastar el comportamiento de la Regresión Logística frente a modelos más densos como *Random Forest y XGBoost*, el estudio desafía empíricamente la supuesta superioridad de los métodos de ensamble sobre las aproximaciones estadísticas tradicionales. A la par, el trabajo examina el efecto de la técnica SMOTE+Tomek sobre la métrica final, arrojando luz sobre qué tan efectivas resultan las alteraciones sintéticas de remuestreo cuando el algoritmo intenta equilibrar el panorama entre empresas exitosas y negocios en riesgo.

De manera complementaria, la incorporación de técnicas de agrupamiento como *K-Means* y DBSCAN ayuda a reconocer perfiles financieros diferenciados dentro del segmento empresarial objeto de estudio. Esta aproximación trasciende la lógica binaria de clasificación y facilita una comprensión más profunda de la heterogeneidad existente entre pequeñas empresas ecuatorianas.

La practicidad del estudio se refleja en su capacidad para apoyar procesos de evaluación crediticia, diseño de programas de fortalecimiento empresarial, formulación de políticas públicas y toma de decisiones estratégicas dentro de las propias empresas. La posibilidad de anticipar un desempeño financiero permite actuar de manera preventiva, reduciendo la

dependencia a decisiones que se toman cuando el problema de insolvencia o fracaso empresarial ya se han materializado.

Finalmente, la investigación se enmarca en un esfuerzo orientado a consolidar el uso de técnicas de ciencia de datos aplicadas al análisis de fenómenos económicos nacionales. La disponibilidad de información financiera oficial, combinada con metodologías reproducibles e interpretables, ofrece una oportunidad para generar evidencia empírica, evitando la simple transferencia de modelos desarrollados en otras realidades económicas.

# CAPÍTULO 2

# 2. REVISIÓN DE LITERATURA

## 2.1. Estado del arte

### *2.1.1.* Predicción del desempeño financiero empresarial

La predicción del desempeño financiero es una de las líneas de investigación reconocida ampliamente en el campo de las finanzas corporativas (Basali, 2025; Zainudin et al., 2024). El interés por esta temática nace de la necesidad de identificar de manera oportuna escenarios de insolvencia o estrés financiero y facilitar la toma de decisiones estratégicas orientadas a preservar la sostenibilidad de la empresa. Bajo este enfoque, los estados financieros han sido considerados una fuente de información privilegiada que es capaz de revelar patrones asociados al comportamiento futuro de las empresas (Burja, 2011).

Los primeros trabajos en este campo de conocimiento estuvieron orientados a la predicción del fracaso empresarial. Uno de los estudios más reconocidos y citados por la literatura financiera fue el desarrollado por Altman (1968), quien propuso el modelo Z-score basado en análisis discriminante múltiple. Sus resultados demostraron que ciertas razones financieras permitían categorizar a las empresas según su probabilidad de quiebra con niveles de precisión aceptables. Este trabajo fue un punto de inflexión al evidenciar que la información contable no solo describía la situación financiera presente, sino que también contenía señales útiles para anticipar eventos futuros.

Posteriormente, Ohlson (1980) introdujo modelos de regresión logística. A diferencia del enfoque propuesto por Altman, el modelo de regresión logística permitió estimar probabilidades individuales de ocurrencia del evento de interés sin depender de supuestos estrictos sobre normalidad multivariante. Este importante avance permitió consolidar el uso de modelos estadísticos para el análisis predictivo de fenómenos financieros y abrió nuevas posibilidades para la evaluación del riesgo financiero.

En los años posteriores, la literatura financiera evolucionó desde la predicción de fracaso empresarial hacia el estudio del desempeño financiero futuro. Este cambio respondió a la necesidad de adoptar enfoques preventivos capaces de identificar factores asociados a la sostenibilidad empresarial antes de que las empresas experimentaran escenarios de insolvencia financiera y posterior fracaso empresarial. Estudios posteriores confirmaron que las ratios financieros relacionados con rentabilidad, eficiencia endeudamiento, liquidez y estructura de capital poseen capacidad predictiva respecto al comportamiento futuro de las organizaciones (Romero Espinosa, 2013; Sánchez Pacheco et al., 2025; Zambrano-Farías, Sánchez-Pacheco, Martínez Mayorga, et al., 2022).

No obstante, gran parte de la evidencia disponible ha sido generada en economías desarrolladas, caracterizadas por mercados financieros más dinámicos y organizados, empresas diferenciadas con mayores niveles de formalización organizacional. Esta situación no permite

la generalización de los resultados hacia economías emergentes como la de Ecuador, donde las pequeñas empresas enfrentan restricciones relacionadas con acceso al crédito, menor capitalización y elevada exposición a cambios del entorno económico.

### *2.1.2.* Aprendizaje automático aplicado a las finanzas

El desarrollo acelerado de la capacidad de almacenamiento y procesamiento de información ha motivado la incorporación de técnicas de *machine learning* en el análisis financiero. Al comparar con los enfoques estadísticos tradicionales, estos algoritmos tienen la capacidad de identificar relaciones complejas, interacciones no lineales y patrones ocultos presentes en el *big data* (Hastie et al., 2017).

Uno de los modelos de aprendizaje supervisados utilizados con frecuencia es la Regresión Logística, cuya aplicación y uso en la literatura financiera especializada se justifica por su fortaleza en la aplicación estadística, facilidad interpretativa y capacidad para estimar probabilidades asociadas a eventos financieros específicos (Hosmer et al., 2013). A pesar de la aparición de nuevos modelos que son más sofisticados, continúa siendo un referente metodológico para evaluar el desempeño de técnicas predictivas.

Posteriormente, los métodos de ensamble adquirieron protagonismo debido a su capacidad para mejorar la precisión predictiva. La técnica de *Random Forest*, desarrollada por Breiman (2001), introdujo una estrategia basada en la combinación de múltiples árboles de decisión construidos sobre muestras *bootstrap* y subconjuntos aleatorios de variables. Este procedimiento permite reducir la varianza de las predicciones y mejorar la capacidad de generalización del modelo.

Más recientemente, *XGBoost* se consolidó como una de las herramientas más utilizadas en competencias de *machine learning* y aplicaciones empresariales. Su mecanismo de *gradient boosting* incorpora regularización, optimización eficiente y estrategias para reducir el sobreajuste, alcanzando elevados niveles de precisión en problemas complejos de clasificación y regresión (Chen & Guestrin, 2016).

Sin embargo, la reciente literatura no presenta un consenso respecto a la superioridad de los algoritmos complejos frente a los tradicionales modelos lineales. Mientras algunos estudios reportan considerables mejoras utilizando técnicas de ensamble, otros concluyen que dichas diferencias son marginales cuando las relaciones subyacentes presentan estructuras relativamente simples o cuando la interpretabilidad constituye un requisito fundamental (Durica et al., 2019; Pietrzak, 2022). Esta discrepancia se justifica con la comparación empírica de distintos algoritmos en contextos específicos como el ecuatoriano.

### *2.1.3.* Desbalance de clases y técnicas de remuestreo

Uno de los desafíos metodológicos identificados con más frecuencia en problemas de clasificación financiera corresponde al desequilibrio o desbalance entre categorías. Este

fenómeno ocurre cuando una de las clases posee una frecuencia considerablemente superior a la otra, generando que los modelos tiendan a favorecer la categoría mayoritaria y disminuyan su capacidad para identificar correctamente las observaciones menos frecuentes (He & Garcia, 2009).

Con el objetivo de disminuir esta limitación, Chawla et al. (2002) desarrollaron la técnica SMOTE (*Synthetic Minority Over-sampling Technique*), su finalidad es generar observaciones sintéticas de la clase minoritaria mediante interpolación entre vecinos cercanos. Este procedimiento aumenta la representatividad de la categoría menos frecuente sin recurrir a la duplicación de observaciones existentes.

Asimismo, Tomek (1976) propuso un método orientado a identificar y eliminar pares de observaciones ambiguas ubicadas en la frontera entre clases. La combinación de ambos procedimientos dio origen a la estrategia híbrida SMOTE+Tomek, ampliamente utilizada para mejorar el equilibrio del conjunto de entrenamiento y reducir la presencia de ruido en los datos.

A pesar de su popularidad, la evidencia actual sugiere que los beneficios del remuestreo dependen del algoritmo implementado, de la magnitud del desbalance y de las características particulares del conjunto de datos analizado. Por tanto, la efectividad de estas técnicas debe evaluarse empíricamente antes de asumir mejoras generalizadas del desempeño predictivo.

### *2.1.4.* Segmentación empresarial mediante aprendizaje no supervisado

Las pequeñas empresas se suelen analizar como un segmento de iguales características debido a los criterios utilizados para su categorización. Sin embargo, investigaciones recientes han demostrado que las empresas que pertenecen a una misma categoría pueden presentar considerables diferencias en términos de estructura económica y financiera, capacidad operativa, rentabilidad y estrategias de crecimiento.

Las técnicas de agrupamiento permiten identificar perfiles semejantes dentro de poblaciones diversas, facilitando la comprensión de estructuras latentes presentes en los datos. Entre ellas, *K-Means* es uno de los algoritmos más utilizados debido a su simplicidad conceptual y eficiencia computacional. Su funcionamiento consiste en asignar iterativamente las observaciones a un conjunto predefinido de centroides, minimizando la variabilidad interna de cada conglomerado (MacQueen, 1967).

Por otra parte, DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)* identifica agrupamientos a partir de la densidad local de las observaciones, permitiendo detectar estructuras arbitrarias e identificar valores atípicos sin necesidad de especificar previamente el número de clústeres (Ester et al., 1996). Esta característica resulta especialmente útil en contextos empresariales caracterizados por alta heterogeneidad.

La incorporación de técnicas no supervisadas complementa los enfoques predictivos tradicionales, ya que permite comprender la diversidad interna del segmento empresarial más allá de una clasificación dicotómica. De este modo, la integración de modelos supervisados y procedimientos de segmentación contribuye a una comprensión más completa del desempeño financiero de las pequeñas empresas.

## 2.2. Marco teórico

### *2.2.1.* Pequeñas empresas y desempeño financiero

Las pequeñas empresas son consideradas uno de los principales impulsores de las economías emergentes debido a su contribución a la generación de empleo, la diversificación de la producción y la dinamización de los mercados locales (Stephen & Elvis, 2011; Welsh et al., 2013). En Ecuador, este segmento empresarial representa una parte importante de la estructura empresarial formal y desempeña un rol destacado en la creación de oportunidades de negocio y en la distribución geográfica de la actividad productiva (Zambrano Farías et al., 2021). De acuerdo con la legislación ecuatoriana, las pequeñas empresas se clasifican tomando en cuenta el volumen anual de ventas, el valor de los activos y el número de trabajadores, lo que permite diferenciarlas de las micro, medianas y grandes empresas (INEC, 2020).

A pesar de la importancia que tienen en la economía local, las pequeñas empresas enfrentan mayores restricciones financieras y operativas que las organizaciones de mayor tamaño. La limitada disponibilidad de capital, el acceso restringido a fuentes formales de financiamiento, la poca capacidad para soportar perturbaciones económicas y las dificultades para incorporar innovaciones tecnológicas incrementan su vulnerabilidad frente a entornos económicos cambiantes (INEC, 2017a). Estas características explican la necesidad de entender los factores que explican su desempeño financiero y desarrollar herramientas que permitan anticipar escenarios favorables o desfavorables para su sostenibilidad operativa.

El desempeño financiero se define como la capacidad que tiene una empresa para generar utilidad o rentabilidad mediante el uso eficiente de sus recursos (Besley & Brigham, 2001).

Desde una perspectiva integral, representa el resultado de la interacción entre decisiones operativas, estrategias de inversión, políticas de financiamiento y condiciones del entorno económico (Besley & Brigham, 2001). Normalmente, el desempeño financiero ha sido estudiado mediante indicadores de rentabilidad, liquidez y solvencia; sin embargo, la literatura actual reconoce que constituye un fenómeno de múltiples dimensiones que puede analizarse desde una perspectiva predictiva (Cultrera & Brédart, 2016; Tascón Fernández & Castaño Gutiérrez, 2012).

En esta investigación, el desempeño financiero futuro se entiende como la condición económica que experimentará la empresa en el período t+1 y que puede anticiparse mediante la información financiera disponible en el período t (Altman, 1968; Ohlson, 1980).

Bajo esta concepción, la predicción del desempeño no busca explicar resultados pasados, sino generar evidencia útil para la toma de decisiones estartégicas y fortalecer la sostenibilidad empresarial (Altman, 1968; Ohlson, 1980).

### *2.2.2.* Indicadores financieros como predictores del desempeño futuro

Los estados financieros representan la principal fuente de información para evaluar la situación económica y el desempeño de las compañías (Zambrano Farías et al., 2021). Su utilidad radica en que muestra de forma resumida las consecuencias financieras derivadas de las decisiones operativas y estratégicas adoptadas por la empresa, permitiendo identificar patrones y tendencias para de esta manera anticipar comportamientos futuros (Ross et al., 2010).

Diversas investigaciones han demostrado que los indicadores financieros tienen la capacidad de predicción con respecto al desempeño empresarial futuro (Altman, 1968; Ohlson, 1980). En este sentido, los indicadores utilizados en el presente trabajo se fundamentan tanto en la teoría financiera como en la evidencia empírica disponible (Brealey et al., 2010; Ross et al., 2010).

Si bien cada indicador contable detalla una dimensión particular de la compañía, su verdadero aporte analítico se materializa al evaluarlos como un sistema integrado. El desempeño económico de una organización no responde a un factor solitario, sino que resulta de la convergencia estructural entre la rentabilidad, la eficiencia operativa y la arquitectura de financiamiento. Al procesar estos componentes de forma simultánea, el análisis trasciende la simple descripción estática. Esto facilita la construcción de un perfil financiero integral, revelando los mecanismos subyacentes que determinan la trayectoria y la sostenibilidad del negocio a largo plazo.

En el ámbito de la predicción, la interacción de estas variables adquiere un peso metodológico determinante. Dos corporaciones pueden registrar niveles de utilidad idénticos durante un mismo ejercicio; sin embargo, sus niveles de apalancamiento o la eficiencia en la rotación de sus activos podrían presentar discrepancias profundas. De igual manera, patrimonios que inicialmente exhiben robustez corren el riesgo de sufrir un deterioro sistemático frente al incremento de la carga financiera o la contracción de las ventas. Esta multiplicidad de condiciones demuestra que la estabilidad corporativa es un fenómeno multicausal. Por consiguiente, pretender estimar la salud financiera futura a partir de una métrica aislada constituye una limitación analítica que reduce drásticamente la capacidad del modelo

El margen operativo es uno de los principales indicadores de eficiencia empresarial, ya que mide la capacidad de la organización para transformar sus ingresos en utilidades derivadas de su actividad principal (Brealey et al., 2010; Ross et al., 2010). Un margen operativo alto muestra un adecuado manejo de costos y gastos asociados a la operación del negocio, mientras que niveles cada vez más bajos de este indicador permiten prever deterioros en el desempeño futuro y complicar la operación de la empresa en el mediano plazo (Almaqtari et al., 2019).

El endeudamiento o apalancamiento financiero representa el grado en que la empresa financia su estructura económica con recursos externos (Ross et al., 2010). Aunque niveles moderados de deuda puede potenciar la rentabilidad, niveles desmesurados de deuda incrementan la exposición al riesgo financiero y con ello experimentar escenario de insolvencia financiera (Akinlo, 2012; Hossain, 2021).

El patrimonio constituyen los recursos propios disponibles para realizar las operaciones empresariales (Rahman & Saif, 2021). En este sentido, empresas con estructuras patrimoniales

sólidas suelen disponer de mayor capacidad para enfrentar inconvenientes financieros y potenciar procesos de crecimiento (Wisna et al., 2023).

La rotación de activos mide la capacidad que tiene la empresa de generar ingresos a partir de sus recursos disponibles (Besley & Brigham, 2001). Una mayor rotación indica que se utilizan los activos de una forma más efectiva, lo que se relaciona con mejores resultados económicos (Sehgal et al., 2021).

La carga financiera refleja la proporción de los gastos financieros con respecto a los ingresos generados por la empresa (Van Horne & Wachowicz, 2010), es decir valores elevados pueden reducir la capacidad de inversión y comprometer la rentabilidad futura (Gepp et al., 2010).

Asimismo, estudios previos (García Pérez & Sánchez Ballesta, 2003; Hossain, 2021; Takon & Atseye, 2015) indican que variables propias de la empresa como el tamaño y la antigüedad de la empresa permiten aproximar características asociadas a economías de escala, experiencia acumulada y capacidad de adaptación, factores que pueden influir sobre la sostenibilidad financiera de las organizaciones.

Vistos de manera aislada, los indicadores contables se limitan a ofrecer un retrato numérico de la actividad económica durante un ejercicio fiscal específico; no obstante, su verdadero valor analítico emerge al conjugarlos dentro de un sistema unificado. A diferencia del diagnóstico financiero tradicional (que suele examinar cada ratio por separado), las arquitecturas predictivas procesan la contribución simultánea de todas las variables, logrando desentrañar correlaciones profundas que escapan por completo a los métodos manuales.

Esta ventaja metodológica se vuelve indispensable al procesar bases de datos masivas, donde la alta densidad de registros y la complejidad de las interacciones superan la capacidad de cualquier evaluación convencional. Ante este volumen de información, el aprendizaje automático explota integralmente el contenido de los estados financieros para emitir pronósticos más robustos, fundamentando la toma de decisiones en evidencia cuantitativa rigurosa. Por esta razón, la selección de predictores adoptada en este estudio no obedece exclusivamente a la teoría contable tradicional, sino a la capacidad intrínseca de cada indicador para optimizar el proceso de entrenamiento de los modelos supervisados.

### *2.2.3.* Aprendizaje supervisado

El aprendizaje supervisado constituye una de las principales ramas del *machine learning* y comprende un conjunto de técnicas encaminadas a construir modelos capaces de observar e identificar relaciones entre conjuntos de variables predictoras o explicativas y una variable objetivo determinada previamente (James et al., 2021).

El objetivo del aprendizaje supervisado consiste en aprender patrones y tendencias que están presentes en los datos históricos para así poder realizar predicciones sobre observaciones futuras (James et al., 2021).

En el ámbito financiero, los modelos supervisados han sido utilizados de forma recurrente para la predicción de insolvencia financiera, riesgo crediticio, fracaso y desempeño empresarial (Hosmer et al., 2013; Hastie et al., 2017).

Su principal ventaja radica en la capacidad de transformar grandes volúmenes de información financiera en herramientas de apoyo para la toma de decisiones (Hastie et al., 2017).

Al momento de seleccionar un algoritmo supervisado, la precisión predictiva deja de ser el único factor determinante; entran en juego variables operativas como la naturaleza del set de datos, su volumen total, la exigencia de interpretabilidad y el objetivo último del estudio. Por esta razón, en el terreno de la analítica financiera es común observar que los modelos de alta complejidad coexisten con alternativas más transparentes y directas, ya que estas últimas facilitan la trazabilidad sobre el peso de cada variable en la decisión final. Esta dualidad impulsa un enfoque comparativo integral, donde múltiples arquitecturas se evalúan sobre una misma base empírica para identificar aquella que consiga el equilibrio óptimo entre poder predictivo, estabilidad estadística y legibilidad de resultados.

Bajo este criterio, el contraste analítico trasciende la simple búsqueda del mayor porcentaje de aciertos. Se vuelve imperativo examinar el comportamiento de cada modelo ante retos estructurales específicos, tales como la no linealidad de las relaciones, la alta correlación entre predictores, la asimetría de las distribuciones o el desbalance severo de clases. En consecuencia, poner a prueba de forma simultánea distintas opciones metodológicas dota al investigador de una comprensión profunda sobre las fortalezas y debilidades de cada enfoque, garantizando que la elección final responda de manera idónea a las exigencias particulares del problema financiero analizado.

En esta investigación se implementaron tres algoritmos pertenecientes a diferentes paradigmas de aprendizaje. Por un lado, la regresión logística es un modelo probabilístico citado con frecuencia para resolver problemas de clasificación binaria, siendo su principal fortaleza la interpretabilidad de los resultados, ya que permite estimar cómo las variaciones en los predictores inciden en la probabilidad asociada al evento de interés (Hosmer et al., 2013). A pesar de que existen técnicas más sofisticadas, este algoritmo sigue siendo una referencia para evaluar el desempeño de nuevos modelos.

Por otro lado, la técnica Random Forest, propuesto por Breiman (2001), corresponde a un método de ensamble que se fundamenta en la construcción de múltiples árboles de decisión independientes. La combinación de las predicciones individuales mediante votación mayoritaria favorece a la reducción de la varianza y mejorar la capacidad de generalización del modelo (Breiman, 2001). Una de las ventajas de este modelo es la tolerancia a valores extremos y la estimación de importancia de variables (Bozkurt & Kaya, 2022; Mahmood et al., 2025).

Por su parte, la técnica de aprendizaje supervisado XGBoost representa una de las aplicaciones más eficientes de los algoritmos de *gradient boosting* (Campillo et al., 2018; Romero Martínez et al., 2021). La ejecución de este algoritmo se basa en la construcción secuencial de árboles de decisión que corrigen iterativamente los errores cometidos por modelos previos, incorporando mecanismos de regularización destinados a minimizar el sobreajuste

(Chen & Guestrin, 2016). De esta manera, su elevado nivel de precisión ha favorecido su utilización en múltiples aplicaciones predictivas.

La comparación de estas tres técnicas permite evaluar si modelos de complejidad creciente ofrecen ventajas en la predicción del desempeño financiero futuro de pequeñas empresas ecuatorianas.

Un mayor grado de complejidad computacional en un algoritmo no garantiza, por sí solo, un salto en la capacidad predictiva. La evidencia empírica demuestra que, frente a escenarios con relaciones lineales simples o un volumen de datos limitado, los modelos estadísticos convencionales compiten al mismo nivel que las técnicas más avanzadas. Sin embargo, cuando el problema introduce interacciones intrincadas, dinámicas no lineales o una alta dimensionalidad de variables, los métodos de ensamble despliegan un valor estratégico superior al descifrar patrones que rebasan el alcance de los enfoques lineales.

Esta divergencia justifica la necesidad imperativa de contrastar empíricamente las arquitecturas bajo un mismo entorno de datos. En lugar de sobredimensionar las virtudes teóricas de un algoritmo, el rigor metodológico exige medir su rendimiento frente a las restricciones particulares del objeto de estudio. Para una investigación aplicada como la presente, este enfoque empírico resulta fundamental, ya que otorga una justificación sólida para la selección del modelo final, respaldada de manera directa por las características operativas del contexto analizado.

### *2.2.4.* Interpretabilidad de modelos predictivos

La adopción recurrente de algoritmos complejos en el ámbito financiero ha generado preocupación respecto a la transparencia de los procesos de decisión automatizados. Aunque modelos como Random Forest y XGBoost pueden alcanzar niveles altos de precisión, con frecuencia son considerados como "cajas negras" debido a la dificultad para comprender cómo generan sus predicciones (Molnar, 2022).

La interpretabilidad hace referencia a la capacidad de explicar el comportamiento de un modelo y comprender la contribución relativa de cada variable explicativa sobre las decisiones generadas (Molnar, 2022). En el entorno financiero, esta característica tiene una importancia particular debido a la necesidad de justificar decisiones relacionadas con otorgamiento de crédito, evaluación empresarial o diseño de estrategias de intervención.

Una de las estrategias que se usan con frecuencia para mejorar la interpretabilidad consiste en analizar la importancia relativa de las variables explicativas (Molnar, 2022). Este procedimiento identifica cuáles son los factores que inciden de manera significativa sobre el desempeño del modelo, transformando algoritmos complejos en herramientas útiles para la generación de conocimiento y el apoyo a la toma de decisiones.

En aplicaciones financieras, la interpretabilidad adquiere un papel determinante porque las decisiones derivadas de los modelos predictivos pueden tener implicaciones económicas importantes para empresas, instituciones financieras e inversionistas. Una predicción con alta precisión resulta de utilidad únicamente si también es posible comprender cuáles fueron los

factores que condujeron a dicha estimación. De lo contrario, el proceso de toma de decisiones puede depender de resultados difíciles de justificar desde una perspectiva técnica o administrativa.

Comprender la jerarquía e importancia relativa de las variables otorga visibilidad directa sobre cuáles indicadores financieros gobiernan las estimaciones del modelo. Más allá de servir como un mecanismo de validación técnica, este desglose arroja luz sobre los determinantes reales del comportamiento corporativo a futuro. Por esta razón, la interpretabilidad deja de ser un mero accesorio y se consolida como un pilar complementario del rendimiento predictivo; su aplicación transforma a los algoritmos de caja negra en instrumentos idóneos para respaldar decisiones estratégicas y decodificar con precisión la dinámica financiera evaluada.

### *2.2.5.* Aprendizaje no supervisado y segmentación empresarial

En contraste con el aprendizaje supervisado, el objetivo de las técnicas no supervisadas es identificar patrones presentes en los datos sin necesidad de determinar una variable objetivo previamente definida, siendo su finalidad descubrir estructuras latentes, relaciones ocultas y agrupamientos naturales dentro de conjuntos de observaciones (Hastie et al., 2017).

Dentro de estas técnicas, el algoritmo K-Means es uno de los algoritmos que se usan con frecuencia para la segmentación empresarial. La ejecución de esta técnica se basa en la asignación iterativa de observaciones a un conjunto predefinido de centroides, minimizando la variabilidad interna de cada conglomerado (MacQueen, 1967).

Por otro lado, DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) identifica agrupamientos a partir de la densidad local de las observaciones. A diferencia de *K-Means*, no requiere especificar previamente el número de clústeres y posee la capacidad de detectar observaciones atípicas o ruido dentro del conjunto de datos (Ester et al., 1996).

Adicionalmente, el Análisis de Componentes Principales (PCA) es una técnica que permite reducir la dimensionalidad orientada a mostrar información multidimensional a través un número reducido de componentes que preservan la mayor proporción posible de variabilidad original (Jolliffe & Cadima, 2016). De esta manera, su aplicación facilita la visualización e interpretación de estructuras complejas.

# CAPÍTULO 3

# 3. DESARROLLO

## 3.1. Diseño metodológico

El trabajo de titulación propuesto adoptó un enfoque cuantitativo ya que el fenómeno objeto de estudio puede ser representado mediante variables numéricas provenientes de los estados financieros reportados por las empresas a la Superintendencia de Compañías, Valores y Seguros en el periodo 2020 – 2024. Para el realizar el análisis se utilizó información objetiva, verificable y susceptible de procesamiento digital y computarizado, permitiendo identificar patrones y tendencias, estimar relaciones entre variables y desarrollar modelos predictivos basados en evidencia empírica. De acuerdo con Creswell (2009) este enfoque es pertinente cuando el propósito es contrastar hipótesis mediante herramientas estadísticas y generar inferencias sustentadas en datos estructurados.

En cuanto al alcance, este trabajo es de carácter predictivo. En contraste con investigaciones de tipo descriptivas que están orientadas a caracterizar fenómenos observados, la finalidad principal del estudio fue estimar la probabilidad de que una pequeña empresa ecuatoriana presente un desempeño financiero positivo en el ejercicio fiscal siguiente, utilizando como insumo la información contable y financiera disponible en el período previo.

Con respecto al diseño de la investigación fue no experimental, ya que no fue necesario manipular de forma deliberada las variables seleccionadas para el estudio. La información utilizada provino de los estados financieros reportados oficialmente a la Superintendencia de Compañías, Valores y Seguros.

Asimismo, la investigación utilizó datos de corte longitudinal, puesto que la información analizada corresponde al periodo comprendido entre los años 2020 y 2024. Este horizonte de tiempo permitió identificar el desarrollo de las empresas durante un período marcado por considerables variaciones económicas derivadas de la pandemia por COVID-19 y la posterior recuperación económica. La disponibilidad de información durante el periodo objeto de estudio favoreció la construcción de variables temporales y la estimación del desempeño financiero futuro bajo un esquema de predicción $t + 1$.

Finalmente, este trabajo de titulación es considerada como una investigación aplicada dentro del campo de la ciencia de datos. Este tipo de investigación busca generar soluciones prácticas a problemas reales mediante la integración de técnicas estadísticas, algoritmos de aprendizaje automático y herramientas computacionales para la toma de decisiones basada en datos (Provost & Fawcett, 2013). En este caso, el conocimiento generado se orientó a desarrollar modelos capaces de apoyar procesos de evaluación financiera, gestión del riesgo empresarial y formulación de estrategias preventivas dirigidas al segmento de pequeñas empresas ecuatorianas.

## 3.2. Fuente de información y construcción de la base de datos

La información utilizada en esta investigación provino de la Superintendencia de Compañías, Valores y Seguros del Ecuador (SCVS), organismo responsable de regular y supervisar a las sociedades mercantiles del país. Esta entidad pública periódicamente información financiera reportada por las empresas obligadas a presentar sus estados financieros, constituyéndose en una de las fuentes oficiales más completas para el análisis del comportamiento empresarial ecuatoriano.

El período de estudio comprendió los ejercicios fiscales correspondientes a los años 2020, 2021, 2022, 2023 y 2024. La selección de este horizonte temporal respondió a dos criterios fundamentales. En primer lugar, permitió analizar el desempeño empresarial durante un contexto económico excepcional caracterizado por los efectos derivados de la pandemia de COVID-19 y la posterior fase de recuperación. En segundo lugar, garantizó la disponibilidad de información suficiente para construir la variable objetivo basada en la predicción del desempeño financiero del período siguiente (t+1).

Los criterios de selección para identificar a las pequeñas empresas se realizaron tomando en cuenta las normas establecidas por la legislación ecuatoriana vigente, los cuales están relacionadas con variables como el volumen de ventas anuales y el número de trabajadores. Posteriormente, se aplicaron filtros orientados a garantizar la consistencia de la información y la comparabilidad entre observaciones.

La construcción de la base de datos siguió un procedimiento sistemático:

1. Descarga de archivos desde la página oficial de la SCVS.
2. Unificación de los archivos en un único conjunto de datos.
3. Depuración de registros duplicados.
4. Exclusión de empresas que no cumplían con los criterios de clasificación de pequeña empresa.
5. Revisión de inconsistencias contables y validación de variables esenciales para el análisis predictivo.

Finalmente, se verificó la integridad de la información financiera requerida para el cálculo de los indicadores utilizados en el estudio. Las observaciones que presentaban ausencia de datos críticos o inconsistencias imposibles de corregir fueron excluidas del análisis, con el propósito de preservar la calidad del conjunto final de datos.

La fase de depuración y estructuración de la base de datos constituyó un cimiento fundamental para asegurar la solidez de las estimaciones posteriores. En el ámbito del aprendizaje automático, la calidad de los insumos condiciona de manera directa el desempeño algorítmico, dictando su aptitud para descubrir patrones latentes y formular pronósticos estables. Por este motivo, el diseño metodológico exigió un control exhaustivo previo al modelado, verificando la consistencia cruzada entre los reportes contables y la disponibilidad íntegra de los predictores requeridos.

Durante este proceso también se procuró mantener un equilibrio entre la cantidad de información analizada y la calidad del conjunto de datos final. Si bien la eliminación de registros incompletos reduce el número de observaciones disponibles, conservar información

inconsistente podría introducir sesgos durante el entrenamiento de los modelos y afectar la interpretación de los resultados. En consecuencia, los criterios de depuración aplicados buscaron preservar únicamente aquellas observaciones que reunían las condiciones necesarias para representar adecuadamente el comportamiento financiero de las pequeñas empresas analizadas.

**Tabla 1**

*Proceso de construcción y depuración de la base de datos*

| Etapa | Descripción del procedimiento |
|---|---|
| Descarga de información | Obtención de bases financieras anuales publicadas por la SCVS correspondientes al período 2020–2024. |
| Integración de bases | Unión de los archivos anuales para conformar una base longitudinal empresa-año. |
| Identificación del segmento | Selección de empresas clasificadas como pequeñas según los criterios normativos vigentes. |
| Eliminación de duplicados | Depuración de registros repetidos o inconsistentes. |
| Validación contable | Verificación de coherencia entre variables financieras fundamentales. |
| Exclusión de registros incompletos | Eliminación de observaciones sin información indispensable para el análisis. |
| Construcción de indicadores | Cálculo de variables financieras y estructurales utilizadas como predictores. |
| Generación de la variable objetivo | Definición del desempeño financiero futuro para la predicción t+1. |
| Base analítica final | Obtención del conjunto definitivo utilizado en los modelos supervisados y no supervisados. |

El procedimiento descrito evidencia que la preparación de la información constituyó un proceso secuencial, donde cada etapa dependió de la correcta ejecución de la anterior. La integración de las bases permitió consolidar la información histórica de las empresas, mientras que las actividades de depuración y validación aseguraron que los indicadores financieros fueran calculados sobre registros consistentes. Finalmente, la construcción de la variable objetivo y de las variables predictoras dio origen al conjunto analítico utilizado para el entrenamiento de los modelos supervisados y para los procedimientos de segmentación empresarial desarrollados en los capítulos posteriores

## 3.3. Definición de variables

La construcción del conjunto analítico requirió la definición de una variable objetivo asociada al desempeño financiero futuro y un conjunto de variables predictoras derivadas de indicadores financieros y características estructurales de las pequeñas empresas ecuatorianas. La selección de estas variables se sustentó tanto en evidencia empírica previa como en criterios de interpretabilidad y disponibilidad de información oficial.

### *3.3.1.* Variable objetivo

En consonancia con estudios previos (Abou Elseoud et al., 2020; Nimer et al., 2015; Rahman & Saif, 2021), la variable objetivo del estudio correspondió a la rentabilidad futura de la empresa, construida bajo un esquema temporal $t + 1$. Su finalidad consistió en representar la condición financiera que experimentaría la organización en el ejercicio fiscal siguiente utilizando como predictores la información disponible en el período t.

En términos operativos, se definió una variable dicotómica denominada desempeño financiero futuro, codificada de la siguiente manera:

- Valor igual a uno (1): empresa clasificada como rentable en el período $t + 1$.
- Valor igual a cero (0): empresa clasificada como no rentable en el período $t + 1$.

Esta definición permitió formular el problema como una tarea de clasificación binaria susceptible de ser abordada mediante técnicas de aprendizaje supervisado.

### *3.3.2.* Variables predictoras

El conjunto de predictores combinó indicadores financieros con características estructurales de las empresas. Esta selección permitió representar aspectos vinculados con la rentabilidad, el endeudamiento, la operación y el tamaño empresarial, siguiendo antecedentes documentados en estudios financieros previos (Aman, 2019; Gołaś, 2020; Nurhayati et al., 2017; Parrales Choez et al., 2024). Las variables que presentaron distribuciones altamente asimétricas fueron transformadas posteriormente mediante logaritmos.

La selección de los predictores también consideró criterios relacionados con la disponibilidad y la calidad de la información reportada por la SCVS. Aunque existen otras variables potencialmente relevantes para explicar el desempeño financiero, muchas de ellas no se encontraban disponibles de manera homogénea para todas las empresas o presentaban niveles elevados de datos faltantes. En consecuencia, se priorizaron indicadores que, además de contar con respaldo en la literatura especializada, pudieran calcularse de forma consistente para la mayor parte de las observaciones incluidas en el estudio.

Otro aspecto considerado fue la interpretabilidad de las variables. Dado que uno de los objetivos de la investigación consistió en identificar los factores asociados al desempeño

financiero futuro, se optó por incorporar indicadores ampliamente utilizados en el análisis financiero tradicional. Esta decisión facilita que los resultados obtenidos mediante los modelos de aprendizaje automático puedan interpretarse desde una perspectiva financiera y compararse con investigaciones desarrolladas en otros contextos.

La selección de las variables predictoras se sustentó en la literatura especializada sobre predicción del desempeño financiero y fracaso empresarial. En particular, los indicadores de rentabilidad, endeudamiento, patrimonio, rotación de activos y carga financiera fueron seleccionados con base en los aportes de Besley y Brigham (2001), Ross et al. (2010), Brealey et al. (2010), Almaqtari et al. (2019), Rahman y Saif (2021), Hossain (2021), Wisna et al. (2023) y Sehgal et al. (2021), quienes evidencian su capacidad para explicar el desempeño financiero futuro de las organizaciones. Asimismo, las variables estructurales relacionadas con el tamaño y la antigüedad empresarial fueron incorporadas considerando la evidencia presentada por García Pérez y Sánchez Ballesta (2003), Takon y Atseye (2015) y Hossain (2021).

**Tabla 2**

*Variables utilizadas en el estudio*

| Variable | Tipo | Descripción | Literatura previa |
|---|---|---|---|
| Desempeño financiero futuro (t+1) | Objetivo | Clasificación binaria del desempeño financiero del período siguiente. | Abou Elseoud et al. (2020); Ahmeti & Balaj (2023); Shahnia et al. (2020) |
| margen_operativo | Predictora | Capacidad de generar utilidad operativa sobre las ventas. | Durrah et al. (2016); Rahman & Saif (2021); Rao et al. (2025) |
| patrimonio | Predictora | Recursos propios aportados por los propietarios de la empresa. | Singapurwoko & El-Wahid (2011); Balasubramanian et al. (2019) |
| deuda_total | Predictora | Obligaciones financieras totales registradas. | Aman (2019); Nindita (2014); Alarussi & Alhaderi (2018) |
| ratio_endeudamiento | Predictora | Nivel relativo de endeudamiento respecto de los activos. | Vatavu (2014); (Ho & Mohd-Raff (2019) |
| carga_financiera | Predictora | Proporción de gastos financieros respecto de los ingresos. | Gepp et al. (2010) |
| rotacion_activos | Predictora | Eficiencia en la utilización de activos para generar ventas. | Jamali (2012); Abdel-Aziz & Alrabba (2023); Alarussi & Alhaderi (2018) |

| Variable | Tipo | Descripción | Literatura previa |
|---|---|---|---|
| gastos_admin_ventas | Predictora | Participación de gastos administrativos y de ventas. | Baños-Caballero et al. (2014) |
| costos_ventas_prod | Predictora | Relación entre costos de producción y actividad comercial. | Shin (2008) |
| total_gastos | Predictora | Magnitud total de egresos operativos. | Bayaraa (2017); |
| gastos_financieros | Predictora | Costos derivados del financiamiento empresarial. | Dia et al. (2022) |
| log_activos | Predictora | Transformación logarítmica del total de activos. | Takon & Atseye (2015); Steinerowska-Streb (2012); Acquah et al. (2022) |
| log_ingresos | Predictora | Transformación logarítmica de ingresos por ventas. | Quispe-Fernández & Ayaviri-Nina (2021) |
| log_empleados | Predictora | Transformación logarítmica del número de empleados. | Garg et al. (2018); Alarussi & Alhaderi (2018) |
| edad_empresa | Predictora | Antigüedad empresarial medida en años. | Lee (2014); Zhao et al., (2015) |

### *3.3.3.* Fórmulas financieras

Los indicadores financieros fueron calculados utilizando información proveniente del estado de situación financiera y del estado de resultados reportados por las empresas analizadas. Las ecuaciones empleadas para la construcción de las variables predictoras se presentan a continuación.

**Margen operativo:**

$$Margen\ Operativo = \frac{Utilidad\ Operativa}{Ingresos\ por\ Ventas}$$

donde:

- **Utilidad Operativa**: resultado obtenido antes de gastos financieros e impuestos.
- **Ingresos por Ventas**: ingresos ordinarios generados por la actividad económica de la empresa.

**Ratio de endeudamiento**

$$Ratio\ de\ Endeudamiento = \frac{Pasivo\ Total}{Activos\ Totales}$$

donde:

- **Pasivo Total**: suma de obligaciones corrientes y no corrientes.
- **Activos Totales**: recursos económicos controlados por la empresa.

**Productividad laboral**

$$Productividad = \frac{Ingresos\ por\ Ventas}{Número\ de\ Empleados}$$

donde:

- **Ingresos por Ventas**: ventas netas del período.
- **Número de Empleados**: personal registrado por la empresa durante el ejercicio económico.

*Esta variable permitió medir la eficiencia en la generación de ingresos por trabajador.*

**Liquidez corriente**

$$Liquidez = \frac{Activos\ Corrientes}{Pasivos\ Corrientes}$$

donde:

- **Activos Corrientes**: efectivo y demás activos realizables en el corto plazo.
- **Pasivos Corrientes**: obligaciones exigibles dentro del siguiente período contable.

**Antigüedad empresarial**

$$Edad = A_t - A_{constitucional}$$

donde:

- $A_t$: año de observación.
- $A_{constitucional}$: año de constitución o inicio de operaciones de la empresa.

**Variable objetivo: desempeño financiero futuro**

$$Y_i = \begin{cases} 1, & \text{si } ROA_{t+1} > Mediana(ROA_{t+1}) \\ 0, & \text{si } ROA_{t+1} \leq Mediana(ROA_{t+1}) \end{cases}$$

donde:

- $ROA_{t+1}$: rendimiento sobre los activos del período siguiente.
- **Mediana(**$ROA_{t+1}$**)**: valor mediano anual utilizado como punto de corte para clasificar a las empresas con alto desempeño financiero.

$$ROA = \frac{Utilidad\ Neta}{Activos\ Totales}$$

donde:

- **Utilidad Neta**: beneficio después de impuestos.
- **Activos Totales**: valor total de los activos de la empresa.

**Transformaciones logarítmicas**

$$Log(X) = \ln(X + 1)$$

donde $X$ corresponde a las variables continuas con elevada asimetría utilizadas en el estudio:

- Activos totales;
- Liquidez corriente;
- Productividad laboral;
- Edad empresarial;
- Número de empleados.

La transformación logarítmica permitió reducir la influencia de valores extremos y aproximar la distribución de estas variables a la normalidad, favoreciendo la estabilidad de las estimaciones del modelo predictivo.

## 3.4. Análisis exploratorio de datos

Antes de entrenar los modelos se examinó la estructura de la base final. Esta revisión permitió conocer la distribución de los indicadores, detectar valores extremos, observar posibles relaciones entre variables y decidir qué transformaciones debían aplicarse antes del modelado.

El proceso de depuración y estructuración de la base de datos representó un pilar crítico para garantizar la confiabilidad analítica de los resultados. Dentro del aprendizaje automático, la precisión de los datos de entrada determina el éxito operativo de los algoritmos, condicionando su capacidad para extraer patrones y proyectar escenarios estables. En consecuencia, la metodología incorporó una fase de validación rigurosa previa al modelado, asegurando la consistencia interna de los estados financieros y la disponibilidad completa de las variables explicativas.

### *3.4.1.* Estadísticos descriptivos

Para cada variable calculamos indicadores clave como la media, la mediana, la varianza, los valores extremos y los cuartiles. Esta comparación permitió ver con claridad las diferencias entre el promedio y el comportamiento real de las empresas, además de mostrar distribuciones con mucha asimetría. Estos resultados sirvieron de guía para decidir cómo tratar los datos atípicos y qué transformaciones aplicar más adelante.

Del mismo modo, este análisis ayudó a notar que las variables financieras tienen escalas, dispersiones y variabilidades muy distintas entre sí. Gracias a esto pudimos justificar la necesidad de aplicar técnicas de normalización y escalamiento, lo cual es fundamental para que los algoritmos funcionen de manera correcta y estable durante el entrenamiento.

### *3.4.2.* Distribuciones de las variables

La distribución de los indicadores financieros se examinó mediante diagramas de caja. Estas gráficas permitieron observar la concentración de los datos, su nivel de dispersión y la presencia de observaciones alejadas del comportamiento general. La revisión fue especialmente útil en variables como patrimonio, gastos, activos e ingresos, cuyas escalas presentaron diferencias importantes.

La revisión gráfica de las distribuciones reveló diferencias claras en la dispersión y asimetría de las variables analizadas. Mientras algunos indicadores mostraron datos bastante concentrados, otros presentaron colas largas con una presencia notable de valores alejados del promedio. Estos hallazgos confirmaron la gran diversidad real de las pequeñas empresas ecuatorianas y justificaron la necesidad de aplicar técnicas de limpieza y normalización para corregir los sesgos antes de entrenar los modelos predictivos.

### *3.4.3.* Valores extremos

La exploración reveló valores considerablemente alejados de la mayoría de las observaciones. En lugar de eliminarlos de manera automática, se conservaron porque correspondían a registros financieros reales de las empresas analizadas. Su exclusión habría reducido la capacidad del conjunto de datos para representar situaciones de pérdida,

endeudamiento o crecimiento atípico.

Para limitar su influencia sobre el entrenamiento, se aplicaron transformaciones logarítmicas a las variables que mostraron mayor asimetría. De esta forma, se mantuvo la información original sin permitir que unas pocas observaciones dominaran el comportamiento de los modelos.

### *3.4.4.* Correlaciones entre variables

Mediante el cálculo de una matriz de correlaciones de Pearson examinamos cómo se relacionaban los indicadores financieros y estructurales entre sí. Este análisis sirvió para detectar asociaciones clave, localizar posibles variables redundantes y comprobar si existían correlaciones lo suficientemente altas como para alterar el entrenamiento de los algoritmos. Con base en estos resultados se definió el conjunto final de predictores.

Asimismo, esta revisión de correlaciones ayudó a entender la estructura general de las relaciones entre las variables antes de arrancar con el modelado. Aunque una correlación elevada no siempre genera fallas metodológicas, detectarla a tiempo resulta muy útil para prevenir la redundancia entre variables, sobre todo en modelos lineales como la Regresión Logística, donde la multicolinealidad suele desestabilizar las estimaciones y complicar la lectura de los coeficientes.

En contraste, los algoritmos basados en árboles de decisión, como *Random Forest* y *XGBoost*, presentan una mayor capacidad para manejar variables correlacionadas sin que ello comprometa significativamente su desempeño predictivo. No obstante, conocer previamente el grado de asociación entre los indicadores permitió interpretar con mayor criterio los resultados obtenidos posteriormente y evaluar si las relaciones observadas coincidían con la importancia atribuida por los modelos supervisados. De esta manera, la matriz de correlaciones constituyó una herramienta exploratoria que complementó tanto la selección de variables como el análisis de los resultados predictivos.

## 3.5. Preparación de datos

Una vez finalizada la exploración, la base fue preparada para el entrenamiento de los algoritmos. Esta etapa incluyó la transformación de variables asimétricas, el escalamiento de los indicadores utilizados en la segmentación, la codificación de la variable objetivo y la separación de los datos en conjuntos de entrenamiento y prueba.

La preparación de los datos constituyó una etapa previa indispensable antes del entrenamiento de los modelos predictivos. Aunque los algoritmos de aprendizaje automático poseen la capacidad de procesar grandes volúmenes de información, su desempeño depende en gran medida de la calidad, consistencia y estructura del conjunto de datos utilizado durante el entrenamiento. Variables con escalas muy diferentes, distribuciones altamente asimétricas o

registros inconsistentes pueden afectar la capacidad de los modelos para identificar relaciones representativas y disminuir su desempeño cuando se enfrentan a nuevas observaciones.

Por esta razón, las transformaciones aplicadas en esta investigación no tuvieron como finalidad modificar el comportamiento financiero de las empresas, sino adecuar la información para que los algoritmos pudieran procesarla de forma más eficiente. Cada procedimiento fue seleccionado considerando las características particulares de la base de datos construida a partir de la información reportada por la SCVS y las recomendaciones metodológicas ampliamente utilizadas en estudios de ciencia de datos aplicada a problemas de clasificación financiera.

### *3.5.1.* Transformaciones logarítmicas

Debido a la elevada asimetría observada en variables relacionadas con tamaño empresarial, se aplicaron transformaciones logarítmicas naturales a los activos totales, ingresos por ventas y número de empleados. Esta estrategia permitió reducir la influencia de observaciones extremadamente grandes y mejorar la estabilidad numérica de los algoritmos.

La aplicación de transformaciones logarítmicas también permitió reducir la influencia de observaciones con magnitudes considerablemente superiores al resto de la muestra. En información financiera es frecuente encontrar empresas cuyos activos, patrimonio o ingresos difieren ampliamente respecto de otras organizaciones pertenecientes al mismo segmento. Si estas diferencias no se atenúan, algunos algoritmos pueden concentrar el proceso de aprendizaje en dichas observaciones, disminuyendo su capacidad para representar adecuadamente el comportamiento general del conjunto de datos. Por ello, la transformación aplicada buscó mejorar la estabilidad numérica de los modelos sin alterar la interpretación económica de las variables analizadas.

### *3.5.2.* Escalamiento de variables

Las variables utilizadas en los procedimientos de segmentación empresarial fueron sometidas a procesos de estandarización mediante la transformación Z-score, con el propósito de evitar que las diferencias en las unidades de medida y magnitudes de las variables influyeran de manera desproporcionada en el cálculo de distancias empleado por los algoritmos de agrupamiento. Este procedimiento permitió que cada variable aportara de forma equivalente al proceso de formación de conglomerados, especialmente en métodos sensibles a la distancia, como K-Means.

$$Z = \frac{X - \mu}{\sigma}$$

donde:

- $Z$: valor estandarizado de la observación;
- $X$: valor original de la variable;

- $\mu$: media aritmética de la variable;
- $\sigma$: desviación estándar de la variable.

La aplicación de esta transformación genera variables con una media igual a cero y una desviación estándar igual a uno, eliminando el efecto de las diferencias de escala entre indicadores financieros y estructurales. De esta manera, variables como activos totales, productividad, liquidez o número de empleados contribuyen de forma equilibrada al cálculo de las distancias euclidianas utilizadas en los análisis de segmentación empresarial.

En términos interpretativos, un valor $Z = 0$ indica que la observación coincide con la media de la variable; valores positivos reflejan observaciones por encima de la media, mientras que valores negativos representan observaciones inferiores al promedio. Esta estandarización constituyó una etapa previa indispensable para la implementación de los algoritmos K-Means y DBSCAN, garantizando la comparabilidad entre las variables incluidas en el análisis de conglomerados.

### *3.5.3.* Partición entrenamiento-prueba

Para separar la base analítica se establecieron dos conjuntos independientes: uno de entrenamiento y otro de prueba. El primero sirvió para ajustar los parámetros de los modelos, mientras que el segundo se reservó exclusivamente para la fase de evaluación final. Gracias a esta partición, el cálculo de las métricas se realizó sobre empresas que los algoritmos no habían visto previamente, logrando así una estimación mucho más realista de su desempeño predictivo.

Esta división responde a la necesidad de comprobar si los modelos poseen una buena capacidad de generalización. Medir el éxito predictivo usando únicamente los datos de entrenamiento genera una falsa ilusión de precisión, ya que el algoritmo solo demuestra haber memorizado información pasada en lugar de aprender patrones reales. Utilizar datos completamente nuevos permite medir con objetividad cómo se comportará el modelo frente a registros reales que todavía no conoce.

Asimismo, mantener esta barrera estricta previene el sobreajuste. Este problema aparece cuando el modelo memoriza tanto las particularidades del entrenamiento que se vuelve incapaz de interpretar tendencias generales. Evitar este riesgo resulta fundamental en la investigación financiera, donde el objetivo principal no es solo explicar el pasado, sino construir herramientas confiables para la toma de decisiones a futuro.

### *3.5.4.* Codificación de variables

La variable objetivo se almacenó en formato binario, con valores de 0 y 1, para que pudiera ser procesada por los modelos de clasificación. Las etiquetas generadas posteriormente por K-Means y DBSCAN se conservaron como identificadores de grupo y se utilizaron para comparar las características financieras de cada segmento empresarial.

## 3.6. Tratamiento del desbalance

### *3.6.1.* Diagnóstico del desbalance

Uno de los problemas frecuentes en tareas de clasificación financiera es la presencia de distribuciones desiguales entre las categorías de la variable objetivo. Cuando una de las clases posee una frecuencia considerablemente superior a la otra, los algoritmos tienden a favorecer la clase mayoritaria, comprometiendo su capacidad para identificar correctamente los casos menos frecuentes (He & Garcia, 2009).

Por esta razón, previo al entrenamiento de los modelos supervisados, se examinó la distribución de la variable objetivo correspondiente al desempeño financiero futuro. El análisis permitió identificar la existencia de un desbalance moderado entre empresas clasificadas como rentables y no rentables, situación que justificó la evaluación de estrategias de remuestreo.

Con el propósito de determinar el efecto del desbalance sobre el desempeño predictivo, se plantearon dos escenarios analíticos:

- **Escenario 1:** entrenamiento utilizando la distribución original de los datos.
- **Escenario 2:** entrenamiento utilizando datos balanceados mediante técnicas de remuestreo.

### *3.6.2.* Aplicación de SMOTE+Tomek

Para abordar el desbalance se empleó la técnica híbrida SMOTE+Tomek, ampliamente utilizada en problemas de clasificación con clases desiguales.

En una primera fase, el algoritmo SMOTE (*Synthetic Minority Over-sampling Technique*) incrementa la representatividad de la clase minoritaria (empresas con bajo desempeño) sin recurrir a la simple duplicación de registros, lo cual incrementaría el riesgo de sobreajuste.

SMOTE opera en el espacio de características multidimensional seleccionando una observación $x_i$ de la clase minoritaria y determinando sus $k$ vecinos más cercanos mediante el cálculo de la distancia euclidiana:

$$d(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

donde:

- $d(x, y)$: Distancia euclidiana entre la observación $x$ y la observación $y$.
- $n$: Número total de variables o características financieras analizadas.
- $x_j$: Valor de la variable $j$ para la observación $x$.

- $y_j$: Valor de la variable $j$ para la observación $y$.

Una vez identificados los vecinos, el algoritmo genera una nueva observación sintética $x_{nuevo}$ interpolando valores a lo largo del segmento de recta que une a $x_i$ con uno de sus vecinos $x_{zi}$, de acuerdo con la siguiente formulación matemática:

$$x_{nuevo} = x_i + \lambda(x_{zi} - x_i)$$

donde:

- $x_{nuevo}$: Nueva observación sintética generada para la clase minoritaria.
- $x_i$: Observación original seleccionada de la clase minoritaria.
- $x_{zi}$: Uno de los $k$ vecinos más cercanos a la observación $x_i$.
- $\lambda$: Número aleatorio generado uniformemente en el intervalo de 0 a 1.

Posteriormente, los enlaces Tomek eliminan observaciones ambiguas localizadas en la frontera entre clases, mejorando la separabilidad del conjunto de entrenamiento (Tomek, 1976).

Un enlace Tomek se define matemáticamente entre dos observaciones $x$ e $y$ si ambas son sus respectivos vecinos más cercanos. Es decir, no existe ninguna otra observación $z$ en el espacio de datos que cumpla con la condición:

$$d(x,z) < d(x,y) \quad \text{o} \quad d(y,z) < d(x,y)$$

Donde:

- $x$: Observación perteneciente a la clase mayoritaria (empresas rentables).
- $y$: Observación perteneciente a la clase minoritaria (empresas no rentables).
- $z$: Cualquier otra observación dentro del espacio de datos.
- $d(\cdot,\cdot)$: Distancia euclidiana entre dos puntos.

Cuando el algoritmo identifica un enlace Tomek, asume que estas observaciones representan ruido estadístico en la frontera de decisión y elimina la observación perteneciente a la clase mayoritaria, mejorando la separabilidad de las clases.

## 3.7. Modelos supervisados

Con el objetivo de estimar el desempeño financiero futuro, se implementaron tres algoritmos de clasificación supervisada pertenecientes a distintos paradigmas de aprendizaje.

### *3.7.1.* Regresión logística

La regresión logística constituye uno de los modelos probabilísticos más utilizados en problemas de clasificación binaria debido a su simplicidad interpretativa y eficiencia computacional (Hosmer et al., 2013).

Este modelo estima la probabilidad de ocurrencia del evento de interés mediante la función logística:

En este estudio, el evento corresponde a que una empresa presente un alto desempeño financiero futuro, definido como un valor de $ROA_{t+1}$superior a la mediana anual.

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

donde:

- $P(Y = 1 \mid X)$: probabilidad condicional de que la empresa presente un desempeño financiero favorable;
- $Y$: variable dependiente dicotómica, donde $Y = 1$indica alto desempeño financiero futuro y $Y = 0$bajo desempeño financiero futuro;
- $e$: base de los logaritmos naturales ($e \approx 2{,}71828$);
- $\beta_0$: intercepto o término constante del modelo;
- $\beta_1, \beta_2, \ldots, \beta_k$: coeficientes asociados a cada variable explicativa;
- $X_1, X_2, \ldots, X_k$: variables predictoras incluidas en el modelo;
- $k$: número total de variables independientes consideradas.

De manera específica, el modelo estimado en esta investigación incorporó las siguientes variables predictoras:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-x}}$$

Con

$$x = \beta_0 + \beta_1 Log(Activos) + \beta_2 Log(Liquidez) + \beta_3 Endeudamiento + \beta_4 MargenOperativo + \beta_5 Log(Productividad) + \beta_6 Log(Edad) + \beta_7 Log(Empleados)$$

donde:

- $Log(Activos)$: logaritmo natural de los activos totales;
- $Log(Liquidez)$: logaritmo natural del indicador de liquidez corriente;
- $Endeudamiento$: ratio de endeudamiento;
- $MargenOperativo$: margen operativo;
- $Log(Productividad)$: logaritmo natural de la productividad laboral;
- $Log(Edad)$: logaritmo natural de la antigüedad empresarial;
- $Log(Empleados)$: logaritmo natural del número de empleados.

La estimación de los coeficientes se realizó mediante el método de máxima verosimilitud, permitiendo cuantificar el efecto de cada variable sobre la probabilidad de que una empresa alcance un desempeño financiero favorable en el período siguiente. Asimismo, la interpretación de los coeficientes se complementó mediante el cálculo de los odds ratios ($OR = e^{\beta}$), los cuales expresan el cambio relativo en las probabilidades de ocurrencia del evento ante variaciones unitarias de las variables predictoras.

### *3.7.2.* Random Forest

Random Forest es un algoritmo de ensamble basado en múltiples árboles de decisión construidos sobre muestras bootstrap del conjunto de entrenamiento. La predicción final se obtiene mediante votación mayoritaria, lo que contribuye a reducir la varianza y mejorar la estabilidad del modelo (Breiman, 2001).

Entre sus ventajas destacan la capacidad para modelar relaciones no lineales, tolerar valores extremos y estimar la importancia relativa de las variables predictoras.

La construcción de cada árbol de decisión dentro del ensamble no se realiza de forma arbitraria, sino que responde a un proceso de optimización que busca maximizar la pureza de los nodos. En este estudio, donde la variable objetivo es dicotómica, el algoritmo evalúa la calidad de cada partición utilizando la Impureza de Gini:

$$G(m) = 1 - \sum_{i=1}^{K} p_{i,m}^2$$

Donde:

- $G(m)$: Nivel de impureza de Gini para el nodo $m$.
- $K$: Número total de clases en la variable objetivo (en este caso, $K = 2$).
- $p_{i,m}$: Proporción de pequeñas empresas pertenecientes a la clase $i$ dentro del nodo $m$.

La principal ventaja de Random Forest radica en la implementación del mecanismo de *Bootstrap Aggregating* (Bagging). El algoritmo extrae múltiples submuestras aleatorias con reemplazo y construye un árbol independiente sobre cada una, considerando solo un subconjunto aleatorio de variables en cada división. Para clasificar una nueva empresa del conjunto de prueba, el algoritmo recolecta las predicciones y emite un resultado basado en votación mayoritaria:

$$\hat{f}_{rf}(x) = \text{moda}\{\hat{f}_b(x)\}_{b=1}^{B}$$

Donde:

- $\hat{f}_{rf}(x)$: Predicción final del modelo Random Forest para la empresa $x$.
- $B$: Número total de árboles de decisión construidos en el ensamble.
- $\hat{f}_b(x)$: Predicción individual generada por el árbol $b$ para la empresa $x$.

### *3.7.3.* XGBoost

*Extreme Gradient Boosting* (XGBoost) es un algoritmo de ensamble secuencial basada en árboles de decisión optimizados mediante gradiente descendente (Chen & Guestrin, 2016). Su elevada eficiencia computacional y capacidad para capturar interacciones complejas lo han convertido en uno de los algoritmos más utilizados en competencias y aplicaciones reales de aprendizaje automático.

El algoritmo incorpora mecanismos de regularización destinados a reducir el sobreajuste y optimizar el desempeño predictivo.

**Tabla 3**

*Hiperparámetros utilizados en los modelos supervisados*

| Modelo | Hiperparámetro | Valor |
|---|---|---|
| Regresión logística | Penalización | I2 |
| Regresión logística | Solver | lbfgs |
| Random Forest | Número de árboles | 500 |
| Random Forest | Profundidad máxima | 12 |
| Random Forest | Muestras mínimas por hoja | 5 |

| Modelo | Hiperparámetro | Valor |
|---|---|---|
| XGBoost | n_estimators | 300 |
| XGBoost | learning_rate | 0.05 |
| XGBoost | max_depth | 6 |
| XGBoost | subsample | 0.80 |

*Nota*. Elaboración propia a partir de los parámetros definidos en el notebook.

## 3.8. Evaluación del desempeño

La evaluación del desempeño predictivo se realizó utilizando métricas ampliamente aceptadas en problemas de clasificación binaria, considerando tanto la capacidad global de clasificación como el comportamiento específico respecto de cada clase.

### *3.8.1.* Accuracy (exactitud)

Representa la proporción de predicciones correctas respecto del total de observaciones evaluadas.

El indicador Accuracy o exactitud mide la proporción de observaciones correctamente clasificadas por el modelo respecto del total de casos evaluados.

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

donde:

- $VP$: Verdaderos Positivos.
- $VN$: Verdaderos Negativos.
- $FP$: Falsos Positivos.
- $FN$: Falsos Negativos.

### *3.8.2.* Precisión (precisión o valor predictivo positivo)

La Precisión o precisión mide la proporción de predicciones positivas correctas entre todas las observaciones que el modelo clasificó como positivas.

$$Precision = \frac{VP}{VP + FP}$$

donde:

- $VP$: Verdaderos Positivos.
- $FP$: Falsos Positivos.

Valores elevados de precisión indican que el modelo genera pocas clasificaciones positivas erróneas.

### *3.8.3.* Recall (sensibilidad o exhaustividad)

El Recall, también denominado sensibilidad o tasa de verdaderos positivos, representa la capacidad del modelo para identificar correctamente los casos pertenecientes a la clase positiva.

$$Recall = \frac{VP}{VP + FN}$$

donde:

- $VP$: Verdaderos Positivos.
- $FN$: Falsos Negativos.

Un alto valor de recall implica que el modelo logra detectar la mayor parte de los casos positivos existentes.

### *3.8.4.* F1-score (medida F1 o índice F1)

El F1-score corresponde a la media armónica entre la precisión (Precisión) y la sensibilidad (Recall), y constituye una métrica especialmente relevante en problemas de clasificación con desbalance de clases, ya que permite equilibrar ambos criterios de evaluación.

$$F1 = 2\,\frac{Precision \times Recall}{Precision + Recall}$$

donde:

- $Precision$: precisión del modelo.
- $Recall$: sensibilidad del modelo.

El valor del F1-score oscila entre 0 y 1, siendo los valores cercanos a 1 indicativos de un mejor desempeño predictivo.

Esta métrica es particularmente útil en este estudio debido a la naturaleza desbalanceada del conjunto de datos, donde una evaluación basada únicamente en la exactitud podría resultar engañosa.

### *3.8.5.* Área bajo la curva ROC (AUC)

El Área Bajo la Curva ROC (AUC, por sus siglas en inglés) evalúa la capacidad discriminativa del modelo para distinguir entre observaciones positivas y negativas, independientemente del umbral de clasificación seleccionado.

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

donde:

- $TPR$(*True Positive Rate*) corresponde a la sensibilidad del modelo:

$$TPR = \frac{VP}{VP + FN}$$

- $FPR$(*False Positive Rate*) representa la proporción de negativos incorrectamente clasificados como positivos:

$$FPR = \frac{FP}{FP + VN}$$

donde:

- $VP$: Verdaderos Positivos.
- $VN$: Verdaderos Negativos.
- $FP$: Falsos Positivos.
- $FN$: Falsos Negativos.
- $TPR$: Tasa de verdaderos positivos.
- $FPR$: Tasa de falsos positivos.

Los valores del AUC varían entre 0 y 1. Un valor de 0,5 indica una capacidad discriminativa equivalente al azar, mientras que valores superiores a 0,80 reflejan una buena capacidad predictiva del modelo.

## 3.9. Segmentación empresarial

Con el objetivo de identificar perfiles financieros semejantes dentro del conjunto de empresas objeto de estudio, se implementaron técnicas de aprendizaje no supervisado.

### *3.9.1.* K-Means

El algoritmo K-Means es una técnica de aprendizaje no supervisado cuyo objetivo es particionar un conjunto de $N$ observaciones en un número $K$ de conglomerados distintos y no superpuestos.

Este modelo agrupa observaciones minimizando la suma de distancias cuadráticas respecto del centroide asignado (MacQueen, 1967). Esta técnica permitió identificar conglomerados empresariales con características financieras similares.

El funcionamiento matemático del algoritmo se basa en un proceso iterativo de optimización. Su objetivo fundamental es encontrar los centroides de cada grupo de manera que se minimice la varianza interna de cada clúster, conocida métricamente como Inercia o Suma de Cuadrados *Intra-Clúster* (*WCSS*, por sus siglas en inglés). La función objetivo que el algoritmo busca minimizar se define mediante la siguiente ecuación:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

Donde:

- $J$: Función objetivo (Inercia) que cuantifica la dispersión interna total de los clústeres. Un valor menor indica grupos más compactos y homogéneos.
- $K$: Número total de conglomerados a formar (determinado a priori mediante índices de validación).
- $C_k$: Conjunto de pequeñas empresas que han sido asignadas al clúster $k$.
- $x_i$: Vector multidimensional que representa los indicadores financieros estandarizados (Z-scores) de la empresa $i$.
- $\mu_k$: Centroide del clúster $k$, calculado como el vector medio aritmético de todas las empresas asignadas a dicho grupo.
- $||x_i - \mu_k||^2$: Distancia euclidiana al cuadrado entre el perfil financiero de la empresa $x_i$ y el perfil promedio de su clúster $\mu_k$.

### *3.9.2.* DBSCAN

Como método complementario se utilizó DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)*, algoritmo basado en densidad capaz de detectar agrupamientos arbitrarios y distinguir observaciones atípicas sin requerir la especificación previa del número de clústeres (Ester et al., 1996).

La incorporación de DBSCAN permitió contrastar la estabilidad de la estructura identificada mediante K-Means.

A diferencia de K-Means (que asume que todos los grupos tienen formas similares y obliga a clasificar a cada empresa dentro de un grupo), DBSCAN funciona de una manera más flexible frente a la información financiera.

### *3.9.3.* Validación del número óptimo de clústeres

La determinación del número apropiado de conglomerados se realizó utilizando diferentes criterios de validación interna, entre ellos:

- Gap Statistic;
- Silhouette Score;
- Davies-Bouldin Index.

Estas métricas permiten evaluar la calidad de la partición en función de la cohesión interna y la separación entre grupos, proporcionando criterios complementarios para la selección del número de clústeres.

En los algoritmos de agrupamiento no supervisado como K-Means, una de las decisiones más críticas es definir a priori cuántos grupos ($K$) deben formarse, ya que el algoritmo por sí solo no conoce la respuesta correcta. Si se eligen muy pocos grupos, se pasa por alto la diversidad real de los datos; si se eligen demasiados, los grupos se fragmentan y pierden significado. Para resolver esto de forma totalmente objetiva, se utilizan métricas de validación interna que actúan como un "filtro de calidad", evaluando dos principios fundamentales:

1. **La cohesión interna:** Mide qué tan parecidas o compactas son las observaciones que quedan dentro de un mismo grupo. Lo ideal es que los elementos de un clúster compartan características muy cercanas entre sí.
2. **La separación entre grupos:** Mide qué tan distantes o diferentes son los grupos entre ellos. Lo ideal es que el perfil de un conglomerado no se mezcle ni se confunda con el de al lado.

### *3.9.4.* Análisis de componentes principales (PCA)

Con fines de visualización, se aplicó Análisis de Componentes Principales (PCA), técnica que permite proyectar la información multidimensional en un espacio reducido preservando la mayor proporción posible de la variabilidad original. Cuando se trabaja con bases de datos financieras que contienen múltiples indicadores (rentabilidad, liquidez, endeudamiento, tamaño), el espacio de características se vuelve hiperdimensional, lo que imposibilita su representación gráfica directa y dificulta el análisis heurístico (Jolliffe & Cadima, 2016).

El PCA es una técnica de transformación lineal no supervisada que proyecta la información multidimensional original en un nuevo sistema de coordenadas ortogonales (los componentes principales), garantizando que la mayor proporción posible de la variabilidad original se preserve en las primeras dimensiones.

El procedimiento matemático del PCA requiere que la matriz de datos original esté rigurosamente estandarizada (proceso detallado previamente en la sección 3.5.2). A partir de la matriz de datos estandarizada $X$, de dimensión $n \times p$ (donde $n$ es el número de pequeñas empresas y $p$ el número de variables financieras), el primer paso consiste en calcular la matriz de covarianza Σ:

$$\Sigma = \frac{1}{n-1} X^T X$$

Donde:

- Σ: Matriz de covarianza simétrica de dimensión $p \times p$.
- $n$: Número total de observaciones (ej. el total de empresas ecuatorianas en el conjunto de datos).
- $X$: Matriz de datos financieros previamente estandarizados.
- $X^T$: Matriz transpuesta de $X$.

El núcleo del PCA consiste en resolver el problema de valores propios (*eigenvalues*) y vectores propios (*eigenvectors*) asociado a la matriz de covarianza. Matemáticamente, se busca encontrar un vector $v$ y un escalar $\lambda$ que satisfagan la ecuación fundamental:

$$\Sigma v = \lambda v$$

Donde:

- $v$: Vector propio (*eigenvector*) que define la dirección del nuevo eje o Componente Principal en el espacio multidimensional.
- $\lambda$: Valor propio (*eigenvalue*) que cuantifica la magnitud de la varianza explicada por ese vector propio específico.
- Σ: Matriz de covarianza.

Esta técnica se utilizó como herramienta exploratoria para facilitar la representación gráfica de los conglomerados obtenidos mediante el algoritmo de clustering.

Al resolver esta ecuación, se obtienen $p$ pares de valores y vectores propios. Estos se ordenan de forma descendente en función de la magnitud de sus valores propios ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$). La proporción de varianza total explicada por el componente $k$-ésimo se calcula como:

$$Varianza\_Explicada_k = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}$$

Proyectar la información de las pequeñas empresas sobre los dos primeros componentes principales (PC1 y PC2) redujo de forma drástica la complejidad dimensional, facilitando la visualización de la separación entre los clústeres financieros en un plano bidimensional.

Al aislar los vectores con los valores propios más altos, el algoritmo sintetizó las correlaciones entre variables redundantes, como la fuerte relación entre los gastos administrativos y el total de egresos ($r = 0{,}828$). Esto permitió condensar el comportamiento financiero de las PYMES ecuatorianas en ejes ortogonales, manteniendo intacta la información estratégica clave.

## 3.10. Procedimiento general del estudio

El procedimiento metodológico desarrollado en esta investigación fue implementado íntegramente mediante un notebook reproducible en Python. La secuencia analítica ejecutada comprendió las siguientes etapas:

1. Descarga e integración de la información financiera proveniente de la SCVS.
2. Selección de pequeñas empresas y depuración de registros.
3. Construcción de indicadores financieros y de la variable objetivo.
4. Análisis exploratorio de datos.
5. Aplicación de transformaciones y preparación del conjunto analítico.
6. División en conjuntos de entrenamiento y prueba.
7. Diagnóstico del desbalance de clases.
8. Aplicación de SMOTE+Tomek sobre el conjunto de entrenamiento.
9. Entrenamiento de modelos supervisados.
10. Evaluación del desempeño predictivo.
11. Estimación de importancia de variables.
12. Segmentación empresarial mediante técnicas no supervisadas.
13. Interpretación integral de los resultados.

# CAPÍTULO 4

## 4. ANÁLISIS DE RESULTADOS

### 4.1. Exploración descriptiva de las pequeñas empresas ecuatorianas

La base final estuvo conformada por 48.190 observaciones empresa-año, correspondientes a 23.683 empresas únicas clasificadas dentro del segmento de pequeñas empresas ecuatorianas durante el período 2020–2024. El proceso de integración inicial partió de tres fuentes: el directorio empresarial, con 219.541 registros y 25 columnas; la base de ranking financiero, con 1.526.841 registros y 54 columnas; y una base unificada inicial de 1.207.906 registros y 78 columnas. Posteriormente, mediante criterios de filtrado por tamaño empresarial, consistencia financiera y disponibilidad de variables relevantes, se obtuvo la base final utilizada en el modelado predictivo.

En términos de delimitación económica, la muestra incluyó empresas con ventas anuales entre USD 100.000,74 y USD 1.000.000,00, activos entre USD 100.008,87 y USD 749.925,82, y una dotación laboral entre 1 y 49 empleados. Estos criterios permitieron focalizar el análisis en empresas de pequeña escala formalmente registradas y con información financiera suficiente para construir indicadores predictivos.

Los resultados del análisis descriptivo de los datos descritos en la Tabla 4 muestran una estructura financiera con notables diferencias. A pesar de que el promedio de ingresos por ventas fue de USD 387.752,56, la mediana se ubicó en USD 323.870,87, lo que muestra una distribución con asimetría positiva. Una situación similar se observó en el tamaño, gastos administrativos, costos de ventas y deuda total. Esta diferencia entre media y mediana confirma que, aun dentro del segmento de pequeñas empresas, existen unidades económicas con niveles de operación significativamente superiores al comportamiento típico del grupo.

**Tabla 4**

*Estadísticos descriptivos generales*

| Variable | n | Media | Desviación estándar | Mínimo | Mediana | Máximo |
|---|---|---|---|---|---|---|
| *Ingresos por ventas* | 48.190 | 387.752,56 | 230.847,42 | 100.000,74 | 323.870,87 | 1.000.000,00 |
| *Ingresos totales* | 48.190 | 391.898,40 | 234.523,84 | 0,00 | 327.747,67 | 3.086.497,43 |
| *Activos* | 48.190 | 292.627,30 | 155.087,93 | 100.008,87 | 253.968,34 | 749.925,82 |
| *Patrimonio* | 48.190 | 111.376,13 | 122.148,05 | -3.567.654,69 | 83.294,20 | 735.378,45 |
| *Utilidad antes de impuestos* | 48.190 | 28.087,33 | 46.441,95 | 0,00 | 11.943,44 | 742.642,47 |
| *Utilidad del ejercicio* | 48.190 | 30.467,05 | 64.593,46 | -5.272.371,36 | 15.275,44 | 728.201,14 |
| *Utilidad neta* | 48.190 | 24.124,86 | 49.235,29 | -1.706.585,15 | 10.647,56 | 728.344,95 |
| *Impuesto a la renta* | 48.190 | 4.163,76 | 7.235,17 | 0,00 | 1.667,03 | 160.416,11 |
| *Número de empleados* | 48.190 | 7,25 | 6,44 | 1,00 | 5,00 | 49,00 |
| *Gastos financieros* | 48.190 | 4.948,12 | 18.493,38 | 0,00 | 470,20 | 743.217,24 |
| *Gastos administrativos y ventas* | 48.190 | 147.535,24 | 172.817,48 | 0,00 | 89.748,54 | 3.556.513,50 |

| Variable | n | Media | Desviación estándar | Mínimo | Mediana | Máximo |
|---|---|---|---|---|---|---|
| *Depreciaciones* | 48.190 | 5.789,98 | 23.402,41 | 0,00 | 183,86 | 4.317.066,00 |
| *Amortizaciones* | 48.190 | 367,04 | 4.390,49 | 0,00 | 0,00 | 404.446,23 |
| *Costos de ventas y producción* | 48.190 | 151.208,58 | 200.514,80 | -159.774,53 | 76.588,09 | 7.768.826,35 |
| *Deuda total* | 48.190 | 14.775,61 | 40.779,96 | 0,00 | 0,00 | 681.685,02 |
| *Total, de gastos* | 48.190 | 191.116,51 | 187.049,26 | 0,00 | 134.633,35 | 5.380.753,00 |

*Nota*. Elaboración propia a partir de datos financieros reportados a la Superintendencia de Compañías, Valores y Seguros.

El patrimonio presentó una media de USD 111.376,13 y una mediana de USD 83.294,20; sin embargo, el valor mínimo fue negativo en USD -3.567.654,69, lo que denota la existencia de empresas con problemas de pérdida acumulada. De forma similar, la utilidad del ejercicio registró un mínimo de USD -5.272.371,36, lo que demuestra que algunas empresas registraron pérdidas dentro del período analizado. Como se indicó, estos valores extremos no fueron eliminados, ya que representan situaciones financieras reales dentro del mercado ecuatoriano y aportan información relevante para la predicción del desempeño futuro.

La Tabla 5 describe la evolución anual de las medianas muestra relativa estabilidad en los ingresos y activos, aunque con una ligera reducción del tamaño operativo a partir de 2022. La mediana de ingresos por ventas pasó de USD 318.711,21 en 2020 a USD 322.438,00 en 2024, mientras que los activos medianos disminuyeron de USD 260.271,61 a USD 250.073,61 en el mismo período. La utilidad neta mediana se mantuvo positiva durante todos los años, con un valor máximo en 2021 de USD 11.266,21 y un valor de USD 10.201,72 en 2024.

En términos generales, los estadísticos descriptivos permiten observar que las pequeñas empresas ecuatorianas presentan una marcada heterogeneidad financiera. Las diferencias identificadas entre los valores mínimos, máximos y las medidas de tendencia central reflejan que, aun perteneciendo a un mismo segmento empresarial, las organizaciones poseen estructuras económicas y operativas considerablemente distintas. Este comportamiento justifica la utilización de modelos capaces de identificar relaciones complejas entre múltiples variables, ya que asumir un comportamiento homogéneo podría conducir a estimaciones poco representativas.

Asimismo, la dispersión observada en varios indicadores financieros evidencia que el desempeño empresarial responde a múltiples factores que interactúan de manera simultánea. Desde una perspectiva analítica, esta variabilidad constituye una característica esperada en bases de datos financieras reales y representa uno de los principales desafíos para el desarrollo de modelos predictivos robustos. En consecuencia, los resultados descriptivos respaldan la necesidad de aplicar técnicas de aprendizaje automático capaces de adaptarse a este nivel de heterogeneidad.

**Tabla 5**

*Medianas financieras por año*

| Año | Ingresos por ventas | Activos | Patrimonio | Utilidad neta | Empleados | Total de gastos | Deuda total |
|---|---|---|---|---|---|---|---|
| 2020 | 18.711,21 | 260.271,61 | 82.328,16 | 9.408,61 | 6 | 126.537,64 | 0,00 |
| 2021 | 333.281,41 | 260.718,06 | 84.773,23 | 11.266,21 | 6 | 127.480,57 | 0,00 |
| 2022 | 324.013,13 | 253.819,18 | 82.775,36 | 11.007,10 | 5 | 123.051,57 | 0,00 |
| 2023 | 323.872,65 | 251.609,37 | 83.503,42 | 11.160,86 | 4 | 141.724,68 | 0,00 |
| 2024 | 322.438,00 | 250.073,61 | 83.200,54 | 10.201,72 | 4 | 144.281,70 | 0,00 |

Un hallazgo relevante es la reducción progresiva del número mediano de empleados, que pasó de 6 trabajadores en 2020 y 2021 a 4 trabajadores en 2023 y 2024. Este comportamiento puede reflejar ajustes operativos, procesos de racionalización de costos o cambios en la estructura laboral de las pequeñas empresas durante el período posterior a la pandemia.

La evolución observada durante el período de estudio también debe interpretarse considerando el contexto económico en el que operaron las empresas. Los años analizados abarcan tanto las consecuencias derivadas de la pandemia de COVID-19 como el proceso de recuperación registrado posteriormente. Por ello, las variaciones identificadas entre un ejercicio fiscal y otro no necesariamente reflejan cambios estructurales permanentes en las empresas, sino también el efecto de condiciones macroeconómicas excepcionales que influyeron sobre la rentabilidad, el nivel de actividad y la capacidad de recuperación del sector productivo.

La Tabla 6 muestra un análisis descriptivo de la variable edad, se puede apreciar que en promedio la edad de las compañías fue de 10,50 años, con una mediana de 8 años. El 25 % de las observaciones correspondió a empresas con una antigüedad igual o inferior a 3 años, mientras que el 75 % presentó una edad igual o inferior a 15 años. La existencia de empresas con hasta 88 años de antigüedad confirma la presencia de unidades consolidadas dentro del segmento, aunque la mayor parte de la muestra se concentró en empresas relativamente jóvenes.

**Tabla 6**

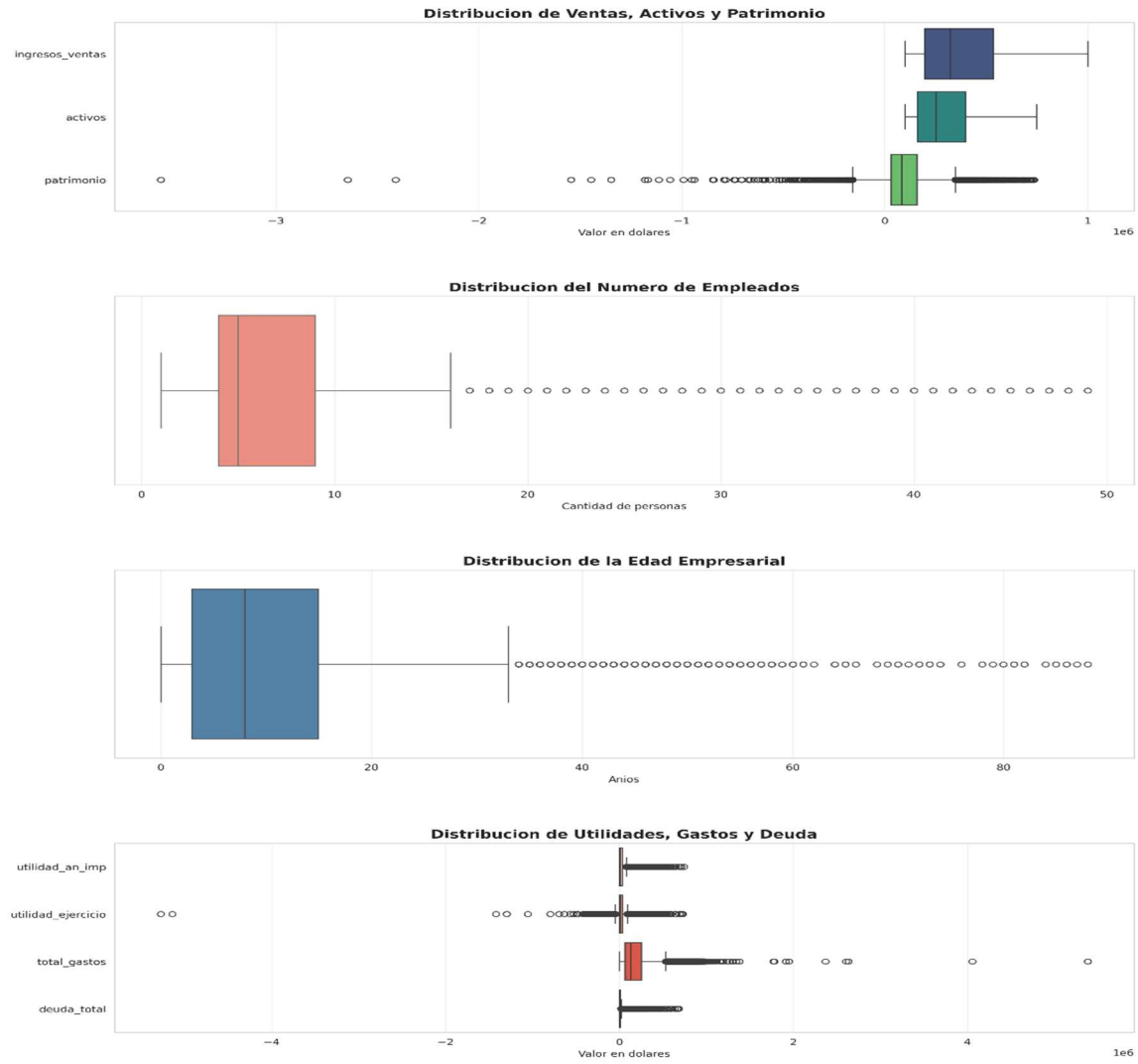
*Estadísticas descriptivas de la edad empresarial*

| Estadístico | Edad empresarial |
|---|---|
| n | 48.190 |
| Media | 10,50 |
| Desviación estándar | 9,29 |
| Mínimo | 0 |
| Percentil 25 | 3 |
| Mediana | 8 |
| Percentil 75 | 15 |
| Máximo | 88 |

La Figura 1 confirma visualmente la presencia de distribuciones asimétricas y valores extremos en varias variables financieras. Las mayores dispersiones se observaron en patrimonio, utilidad del ejercicio, gastos administrativos, costos de ventas, depreciaciones y total de gastos. Este patrón es consistente con la naturaleza heterogénea de las pequeñas empresas ecuatorianas y justifica la aplicación de transformaciones logarítmicas y procedimientos de estandarización en las etapas posteriores del análisis.

**Figura 1**

*Distribución de variables financieras mediante diagramas de caja.*




*Nota*. Fuente: Elaboración propia.

## 4.2. Relaciones entre variables financieras

Con el propósito de identificar el tipo de relación entre las variables financieras seleccionadas, se calculó una matriz de correlación de Pearson (Tabla 7). La correlación más

alta se observó entre las variables gastos administrativos y ventas con el total de gastos ((r = 0,828)). Este resultado es coherente con la literatura financiera, puesto que los gastos administrativos y de ventas forman parte de los esfuerzos que realizan las empresas para generar ingresos.

También se identificó una correlación positiva moderada entre ingresos por ventas y costos de ventas y producción ((r = 0,564)), lo cual refleja que las empresas con mayores niveles de actividad comercial tienden a presentar mayores costos relacionados a la generación de ventas. De manera similar, la relación entre total de gastos e ingresos por ventas fue de (r = 0,528), indicando que el crecimiento operativo suele estar acompañado de un aumento en la estructura de costos y gastos.

La relación entre activos y patrimonio alcanzó un coeficiente de (r = 0,468), lo que sugiere una asociación positiva entre tamaño patrimonial y tamaño económico. Por su parte, la correlación entre patrimonio y utilidad neta fue de (r = 0,400), mostrando que las empresas con mayor base patrimonial tienden a registrar mejores resultados netos, aunque la relación no es lo suficientemente alta como para asumir dependencia perfecta.

**Tabla 7**

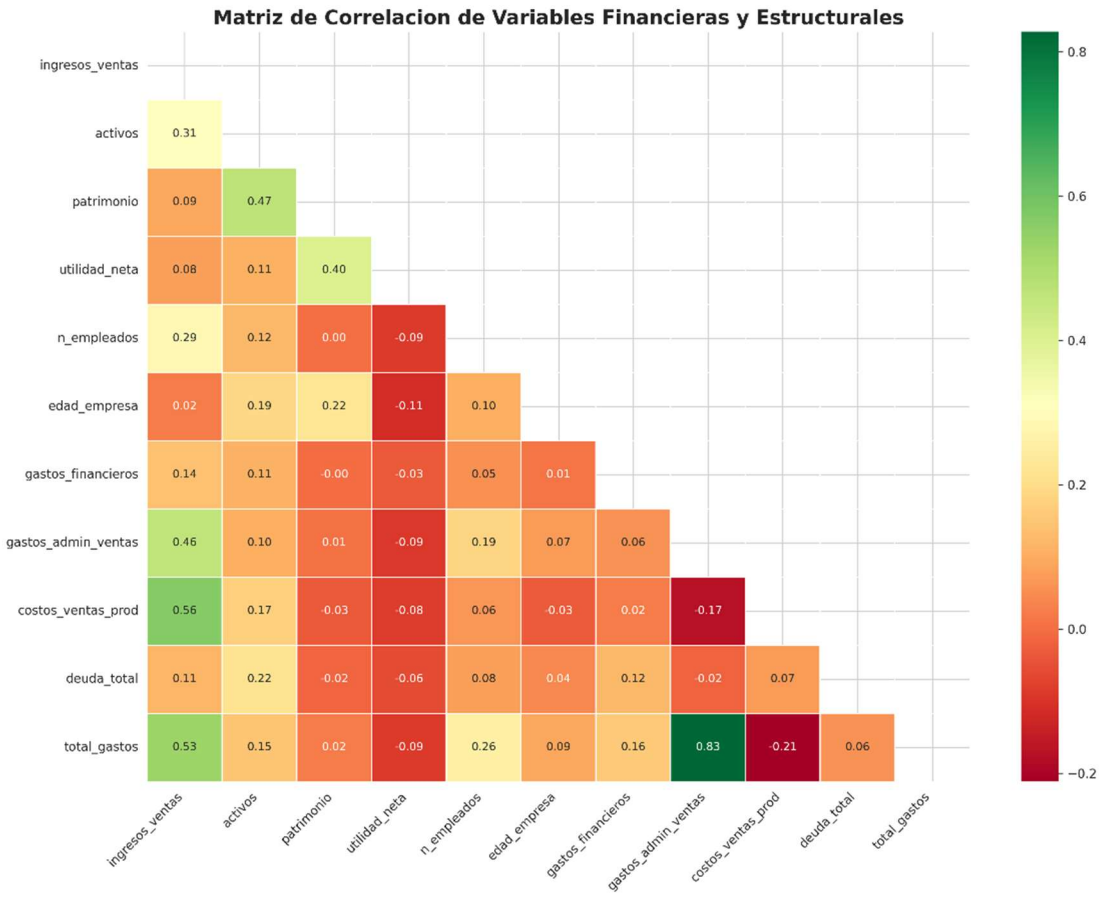
*Principales correlaciones entre variables financieras*

| Variable 1 | Variable 2 | Correlación |
|---|---|---|
| Gastos administrativos y ventas | Total de gastos | 0,828 |
| Ingresos por ventas | Costos de ventas y producción | 0,564 |
| Total de gastos | Ingresos por ventas | 0,528 |
| Activos | Patrimonio | 0,468 |
| Ingresos por ventas | Gastos administrativos y ventas | 0,461 |
| Patrimonio | Utilidad neta | 0,400 |
| Activos | Ingresos por ventas | 0,312 |
| Número de empleados | Ingresos por ventas | 0,287 |

La Figura 2 permite observar que, en general, las correlaciones entre variables se ubicaron en rangos bajos o moderados. No se identificaron patrones generalizados de multicolinealidad severa que pudieran comprometer el entrenamiento de los modelos supervisados. Este hallazgo es relevante porque sugiere que las variables seleccionadas aportan información complementaria para la predicción del desempeño financiero futuro

**Figura 2**

*Matriz de correlación de variables financieras*




*Nota*. Fuente: Elaboración propia.

Los coeficientes obtenidos también ayudaron a definir la estrategia de modelado. Debido a que no se encontraron correlaciones extremadamente altas entre la mayoría de los predictores, se mantuvieron las variables seleccionadas para el entrenamiento de la Regresión Logística. Al mismo tiempo, las relaciones moderadas observadas entre ingresos, costos, gastos, activos y patrimonio justificaron la evaluación de Random Forest y XGBoost, ya que estos algoritmos pueden representar interacciones que un modelo lineal podría no captar. En términos generales, la matriz no evidenció una redundancia suficientemente alta como para excluir variables antes del entrenamiento.

Aunque algunas asociaciones presentan una intensidad moderada, el análisis de correlación demuestra que ninguna variable, por sí sola, explica completamente el desempeño financiero futuro. Este resultado sugiere que el fenómeno estudiado depende de la interacción entre diversos factores financieros y no de un único indicador aislado. En consecuencia, la utilización de algoritmos de aprendizaje automático resulta apropiada, ya que estas técnicas poseen la capacidad de capturar relaciones simultáneas entre múltiples predictores y generar modelos con una mayor capacidad de representación respecto a enfoques basados únicamente en asociaciones bivariadas.

Desde una perspectiva metodológica, estos hallazgos también respaldan la selección del conjunto de variables empleado durante el entrenamiento de los modelos supervisados. La coexistencia de correlaciones de distinta magnitud evidencia que los indicadores aportan

información complementaria, favoreciendo la construcción de modelos capaces de incorporar diferentes dimensiones del desempeño financiero empresarial.

## 4.3. Comparación del desempeño de los modelos predictivos

Uno de los propósitos principales de este estudio consistió en medir la eficacia de diversos algoritmos de aprendizaje automático para pronosticar la situación financiera de las pequeñas empresas en Ecuador. Con este fin, la investigación contrastó el rendimiento de tres modelos supervisados (Regresión Logística, *Random Forest y XGBoost*) bajo dos escenarios operativos: manteniendo la proporción original de la variable objetivo y aplicando una estrategia de remuestreo mediante la técnica SMOTE+Tomek.

Esta estrategia metodológica ayudó a determinar tanto la potencia discriminatoria de cada algoritmo como el impacto real del desbalance de clases en los resultados predictivos.

Evaluar diferentes técnicas de forma paralela va más allá de buscar qué modelo acumula más aciertos; permite entender cómo reacciona cada arquitectura ante retos estructurales típicos de los datos financieros, como la asimetría, las relaciones no lineales o la marcada desigualdad entre categorías. Estudiar estas opciones bajo un mismo marco analítico entrega argumentos sólidos para elegir el modelo que mejor se adapte a la realidad del sector.

Asimismo, esta comparación conjunta aclara si un aumento en la complejidad computacional se traduce en ventajas prácticas reales. Mientras los modelos tradicionales a veces igualan el rendimiento de opciones más densas, las técnicas de ensamble suelen desenterrar patrones que escapan a los enfoques lineales, un factor clave al momento de diseñar herramientas confiables para la toma de decisiones empresariales.

### *4.3.1.* Distribución de la variable objetivo y diagnóstico del desbalance

La construcción de la variable objetivo evidenció un marcado desbalance entre las clases correspondientes al desempeño financiero futuro. En particular, la clase minoritaria representó una proporción considerablemente menor respecto de la clase mayoritaria, situación frecuente en problemas de clasificación financiera y susceptible de introducir sesgos en el entrenamiento de los modelos predictivos. Debido a esta característica, se evaluó la aplicación de la técnica híbrida SMOTE+Tomek con el propósito de mejorar la representación de la clase minoritaria durante el entrenamiento.

El análisis exploratorio mostró que la proporción de empresas clasificadas como rentables superó a la de aquellas que presentaron desempeño desfavorable, situación frecuente en problemas de clasificación financiera. En estos contextos, los algoritmos pueden desarrollar un sesgo hacia la clase mayoritaria, incrementando aparentemente la exactitud global, pero disminuyendo la capacidad para detectar correctamente las observaciones menos frecuentes.

Con el propósito de evaluar la magnitud de este fenómeno, se comparó la distribución de clases antes y después de aplicar SMOTE+Tomek sobre el conjunto de entrenamiento que se muestra en la Tabla 8. La aplicación de SMOTE+Tomek permitió obtener una representación

más equilibrada entre ambas categorías durante la fase de entrenamiento, generando un escenario apropiado para evaluar si la reducción del desbalance se traducía en mejoras efectivas del desempeño predictivo.

**Tabla 8**

*Distribución de clases antes y después del balanceo mediante SMOTE+Tomek*

| Escenario | Clase 0 | Clase 1 | Total |
|---|---|---|---|
| Datos originales | 945 | 16434 | 17379 |
| Datos balanceados | 16071 | 16071 | 32142 |

*Nota*. Fuente: Elaboración propia

Como se observa en la Tabla 8, el conjunto de datos original estuvo conformado por 945 observaciones de la clase 0 y 16 434 de la clase 1, lo que evidencia un marcado desbalance entre ambas categorías. Tras la aplicación de SMOTE+Tomek, el conjunto de entrenamiento alcanzó una distribución equilibrada de 16 071 observaciones por clase. Esta transformación permitió reducir el sesgo asociado a la predominancia de la clase mayoritaria y generar un escenario más adecuado para evaluar si el balanceo contribuía a mejorar la capacidad predictiva de los modelos supervisados.

### *4.3.2.* Resultados utilizando la distribución original de los datos

Los modelos supervisados fueron entrenados inicialmente utilizando la distribución original de la variable objetivo. Los resultados que se obtuvieron mostraron un comportamiento similar entre los tres algoritmos evaluados.

La Regresión Logística alcanzó un área bajo la curva ROC (AUC) de 0,704, mientras que Random Forest obtuvo un valor de 0,701. Por su parte, XGBoost presentó el mejor desempeño bajo este escenario, registrando un AUC de 0,705.

Estos resultados concluyen que la información financiera proveniente de las variables utilizadas posee una capacidad discriminatoria moderada para anticipar el desempeño financiero futuro de las pequeñas empresas ecuatorianas. Asimismo, evidencian que modelos más complejos no necesariamente generan mejoras sustanciales respecto de enfoques probabilísticos tradicionales.

### *4.3.3.* Resultados después del balanceo mediante SMOTE+Tomek

Posteriormente, los modelos fueron reentrenados utilizando el conjunto balanceado mediante SMOTE+Tomek.

Bajo este escenario, la Regresión Logística experimentó una mejora marginal, incrementando su AUC desde 0,704 hasta 0,706. En contraste, los algoritmos de ensamble registraron una disminución en su capacidad discriminatoria.

Random Forest redujo su desempeño hasta un AUC de 0,653, mientras que XGBoost descendió a 0,673.

Estos hallazgos indican que la aplicación de técnicas de remuestreo no produjo beneficios homogéneos sobre todos los algoritmos evaluados. Mientras la regresión logística mostró una ligera mejora, los modelos basados en árboles parecieron verse afectados negativamente por la incorporación de observaciones sintéticas y la eliminación de ejemplos ambiguos.

### *4.3.4.* Comparación integral de los modelos

Con el fin de sintetizar los resultados obtenidos, la Tabla 9 presenta las métricas de desempeño correspondientes a cada algoritmo bajo ambos escenarios analíticos.

**Tabla 9**

*Desempeño de los modelos supervisados para la predicción del desempeño financiero futuro*

| Modelo | Escenario | Accuracy | Precision | Recall | F1 | AUC |
|---|---|---|---|---|---|---|
| Regresion Logistica | Original | 0.9453 | 0.9459 | 0.9993 | 0.9719 | 0.7044 |
| Random Forest | Original | 0.9412 | 0.9564 | 0.9826 | 0.9693 | 0.7013 |
| XGBoost | Original | 0.9454 | 0.9466 | 0.9987 | 0.9719 | 0.7049 |
| Logit_SMOTE | SMOTE_Tomek | 0.4971 | 0.9753 | 0.4803 | 0.6437 | 0.7062 |
| RF_SMOTE | SMOTE_Tomek | 0.8043 | 0.9571 | 0.8303 | 0.8892 | 0.6534 |
| XGB_SMOTE | SMOTE_Tomek | 0.9148 | 0.9534 | 0.9567 | 0.9551 | 0.6734 |

*Nota*. Rxesultados obtenidos en el notebook reproducible.

En este estudio, evaluar métricas como la Precisión o la Sensibilidad (*Recall*) nos ayuda a medir el impacto financiero que tendría equivocarse al clasificar a una empresa.

Si el modelo comete un **Falso Positivo**, significa que predice que una empresa será "rentable" el próximo año, cuando en la realidad va a tener problemas financieros o pérdidas. Para un banco o un inversionista, este es el peor error posible, ya que se traduce en otorgar un crédito que probablemente no se podrá cobrar.

Por otro lado, un **Falso Negativo** ocurre cuando el modelo dice que la empresa "no será rentable", pero en la realidad sí logra tener ganancias. Aunque este error no le hace perder

dinero directamente a un banco, sí representa una oportunidad de negocio desperdiciada por rechazar a un buen cliente.

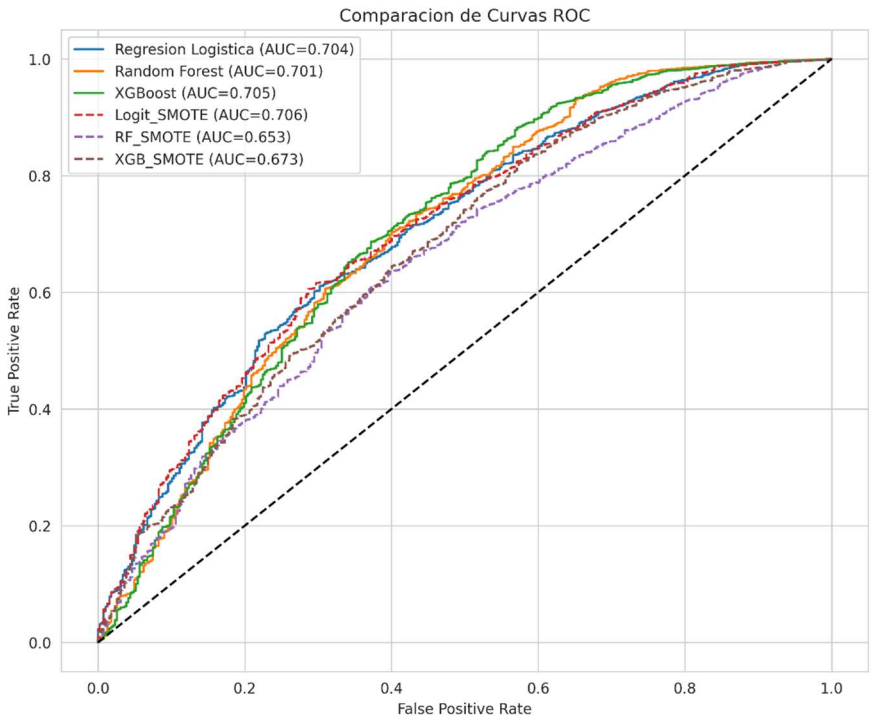
Con los datos originales, el modelo tenía un *Recall* alto (0,9993). Sin embargo, al aplicar la técnica de balanceo SMOTE+Tomek, esta sensibilidad cayó a 0,4803, pero su Precisión subió a 0,9753.

Esto significa que, al balancear los datos, el algoritmo se volvió mucho más conservador. Es decir, el modelo solo clasifica a una empresa como "rentable" cuando está casi completamente seguro de que no va a cometer errores.

La Figura 3 permite visualizar que las diferencias entre los modelos fueron relativamente pequeñas cuando se utilizó la distribución original de los datos. La curva correspondiente a XGBoost presentó el AUC más elevado (0,705), seguida muy de cerca por la Regresión Logística (0,704) y Random Forest (0,701).

**Figura 3**

*Curvas ROC comparativas de los modelos supervisados.*




*Nota*. Fuente: Elaboración propia.

Sin embargo, el escenario balanceado modificó parcialmente este comportamiento. La Regresión Logística logró el mejor desempeño global del estudio con un AUC de 0,706,

mientras que los modelos de ensamble evidenciaron reducciones importantes en su capacidad discriminatoria.

### *4.3.5.* Discusión de resultados en función de las hipótesis

Los hallazgos obtenidos proporcionan información importante para la comprensión del desempeño financiero de las pequeñas empresas ecuatorianas. En relación con la Hipótesis 1, la cual proponía que los modelos de ensamble (*Random Forest y XGBoost*) superarían a la Regresión Logística debido a su capacidad para capturar relaciones no lineales, los resultados no proporcionan evidencia suficiente para respaldarla.

Aunque *XGBoost* obtuvo el mayor AUC en el escenario con datos originales (0,705), la diferencia respecto a la Regresión Logística (0,704) fue mínima y no significativa desde el punto de vista práctico. Tras aplicar SMOTE+Tomek, la Regresión Logística alcanzó el mejor desempeño global (AUC = 0,706), superando a ambos modelos de ensamble. En consecuencia, H1 es rechazada.

Este resultado indica que las relaciones que mantienen los indicadores financieros utilizados y el desempeño financiero futuro presentan un comportamiento lineal o de complejidad moderada, que hace que un modelo probabilístico tradicional alcance niveles de discriminación comparables a los obtenidos mediante algoritmos más sofisticados.

Por otra parte, la Hipótesis 3 postulaba que la aplicación de SMOTE+Tomek mejoraría el desempeño predictivo sobre la clase minoritaria cuando existiera un desbalance estadísticamente relevante. Los resultados muestran un comportamiento diferente, mientras la Regresión Logística experimentó una mejora marginal tras el balanceo, los logaritmos de *Random Forest* y *XGBoost* redujeron significativamente sus valores de AUC. En consecuencia, H3 se acepta parcialmente, dado que los beneficios del remuestreo dependieron del algoritmo empleado y no constituyeron una mejora universal del desempeño predictivo.

En concreto, los resultados obtenidos sugieren que, en el contexto analizado, la predicción del desempeño financiero futuro de las pequeñas empresas ecuatorianas puede realizarse con un poder discriminatorio moderado, alcanzando valores de AUC superiores a 0,70. Asimismo, demuestran que estrategias metodológicas frecuentemente asumidas como deben evaluarse empíricamente en cada contexto específico antes de generalizar su efectividad.

## 4.4. Interpretabilidad de los modelos predictivos

### *4.4.1.* Importancia de variables en Random Forest

La técnica *Random Forest* estima la relevancia de cada predictor a partir de la reducción promedio de impureza generada durante la construcción de los árboles de decisión. Variables

que contribuyen en mayor medida a mejorar la clasificación reciben puntuaciones más elevadas de importancia (Breiman, 2001).

Los resultados muestran que el margen operativo fue la variable con mayor contribución al desempeño del modelo, alcanzando aproximadamente el 21 % de la importancia total. Este hallazgo indica que la capacidad de la empresa para transformar sus ventas en utilidades operativas constituye el principal determinante del desempeño financiero futuro.

La Tabla 10 destaca las variables relacionadas con la estructura patrimonial y el tamaño económico de las empresas, esto es, el patrimonio, el logaritmo de los activos y los gastos administrativos y de ventas.

**Tabla 10**

*Variables más importantes según Random Forest*

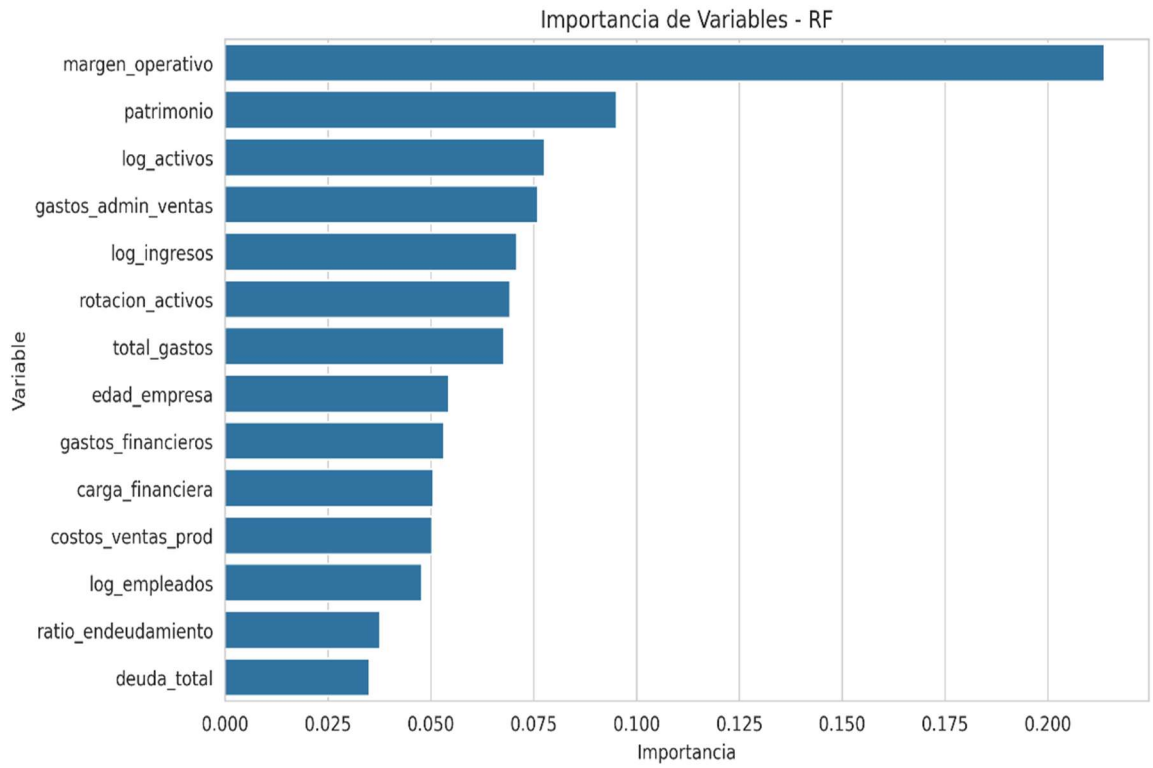
| Variable | Importancia (%) |
|---|---|
| Margen operativo | 21,0 |
| Patrimonio | 9,8 |
| log_activos | 8,7 |
| Gastos administrativos y ventas | 8,1 |
| log_ingresos | 7,5 |
| Total de gastos | 7,1 |
| Costos de ventas y producción | 6,4 |
| Deuda total | 5,9 |
| Número de empleados | 5,1 |
| Edad empresarial | 4,6 |
| Otras variables | 15,8 |

*Nota*. Datos obtenidos del modelo Random Forest implementado en el notebook reproducible.

La Figura 4 evidencia una marcada diferencia entre el margen operativo y el resto de predictores, es decir, el margen operativo explica una proporción sustancial del proceso de clasificación, mientras que las demás contribuyen de manera más distribuida, sugiriendo que el desempeño financiero futuro depende tanto de la eficiencia operativa como de la combinación de múltiples características empresariales.

**Figura 4**

*Importancia de variables en Random Forest.*




*Nota*. Fuente: Elaboración propia.

### *4.4.2.* Importancia de variables en XGBoost

Para *XGBoost* se examinó la participación de cada predictor en las divisiones realizadas por los árboles durante el entrenamiento. Esta medida permitió ordenar las variables según su aporte relativo a las decisiones del modelo. (Chen & Guestrin, 2016).

El margen operativo ocupó nuevamente la primera posición, con una importancia aproximada del 12 %. La coincidencia con Random Forest refuerza la relevancia de este indicador, aunque los porcentajes no deben compararse de forma directa entre ambos modelos porque cada algoritmo calcula la importancia con un procedimiento diferente. Después del margen operativo se ubicaron el patrimonio, la deuda total y el logaritmo de los activos, variables relacionadas con la estructura financiera y el tamaño de la empresa.

**Tabla 11**

*Variables más importantes según XGBoost*

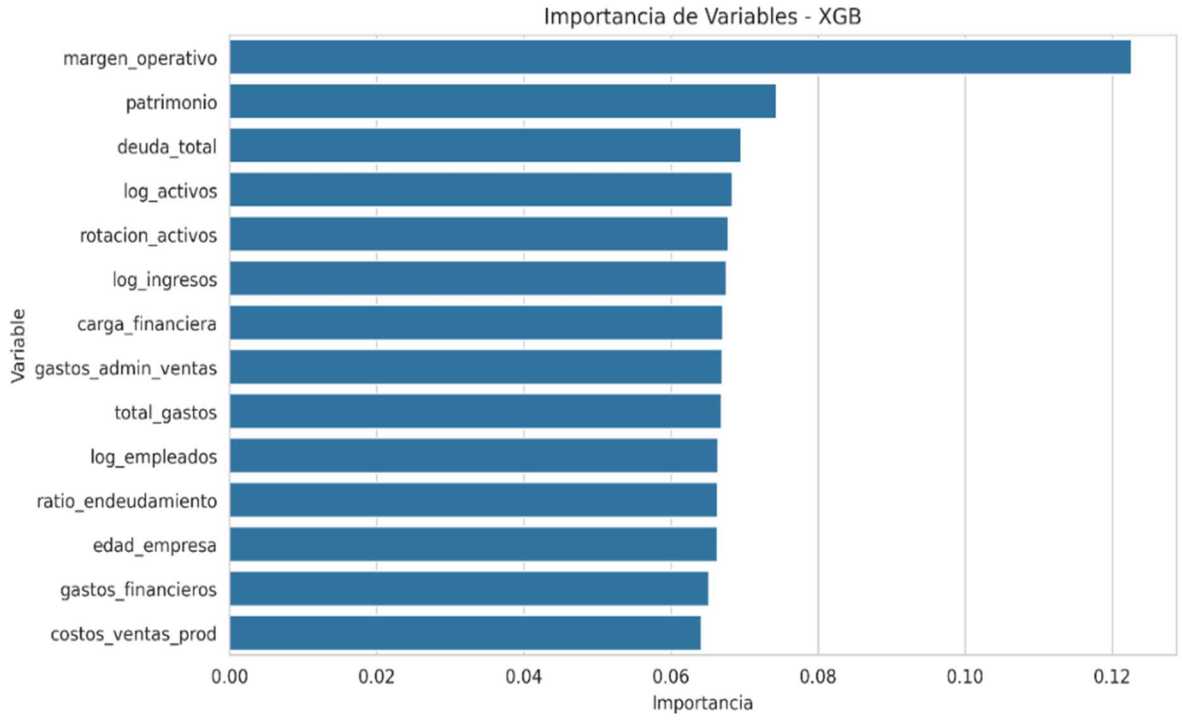
| Variable | Importancia (%) |
|---|---|
| Margen operativo | 12,0 |
| Patrimonio | 10,1 |
| Deuda total | 8,9 |
| log_activos | 8,4 |

| Variable | Importancia (%) |
|---|---|
| Gastos administrativos y ventas | 7,3 |
| Total de gastos | 6,9 |
| log_ingresos | 6,7 |
| Costos de ventas y producción | 6,2 |
| Número de empleados | 5,3 |
| Edad empresarial | 4,7 |
| Otras variables | 23,5 |

*Nota*. Datos obtenidos a partir del modelo XGBoost implementado en Python.

**Figura 5**

*Importancia de variables en XGBoost*




*Nota*. Fuente: Elaboración propia.

A diferencia de Random Forest, XGBoost distribuyó la importancia de manera ligeramente más homogénea entre los predictores. No obstante, la jerarquía general permaneció estable.

### *4.4.3.* Comparación entre ambos enfoques

Al comparar de ambos algoritmos se denota una coincidencia significativa respecto de los determinantes del desempeño financiero futuro. En primer lugar, tanto la técnica de *Random Forest* como la técnica *XGBoost* identificaron al margen operativo como la variable más importante, independientemente de la estrategia de aprendizaje utilizada. Por lo tanto, este resultado refuerza la robustez del hallazgo y disminuye la probabilidad de que la relevancia observada responda a particularidades de un algoritmo específico.

Segundo, ambos modelos resaltan la importancia del patrimonio, los activos y la estructura de gastos, sugiriendo que la capacidad de generar utilidades, mantener una base patrimonial sólida y gestionar eficientemente los costos operativos son factores fundamentales para sostener la rentabilidad futura.

Finalmente, variables tradicionalmente asociadas al tamaño empresarial, como el número de empleados y la antigüedad, presentaron niveles de importancia moderados, lo que indica que, aunque contribuyen al proceso predictivo, su influencia es inferior a la ejercida por los indicadores operativos.

### *4.4.4.* Identificación de los determinantes del desempeño financiero futuro

En términos prácticos, los resultados demuestran que la evolución financiera futura de las pequeñas empresas en Ecuador está condicionada, fundamentalmente, por su habilidad para transformar los ingresos operativos en saldos positivos.

Que el margen operativo se haya consolidado como el predictor clave en los modelos de ensamble representa uno de los descubrimientos más importantes del trabajo. Esto indica que las decisiones sobre control de costos, eficiencia operativa y gestión de gastos impactan de forma mucho más directa en la rentabilidad a futuro que otras variables estructurales de la compañía.

Este resultado refleja fielmente la dinámica del mercado local. En el contexto empresarial del país, una PYME puede generar ventas elevadas o contar con activos importantes; sin embargo, si carece de un control estricto sobre sus costos de producción y gastos administrativos, los recursos se desvanecen antes de convertirse en ganancias reales. Así, la evidencia confirma que asegurar la estabilidad del negocio no depende únicamente de acelerar las ventas, sino de optimizar la administración interna de los recursos.

De este modo, la interpretabilidad de los modelos predictivos cumple una doble función: por un lado, desentraña la lógica interna de los algoritmos y, por el otro, entrega insumos prácticos para diseñar estrategias corporativas enfocadas en blindar la sostenibilidad financiera de las pequeñas empresas.

## 4.5. Segmentación empresarial de las pequeñas empresas ecuatorianas

### *4.5.1.* Determinación del número óptimo de clústeres

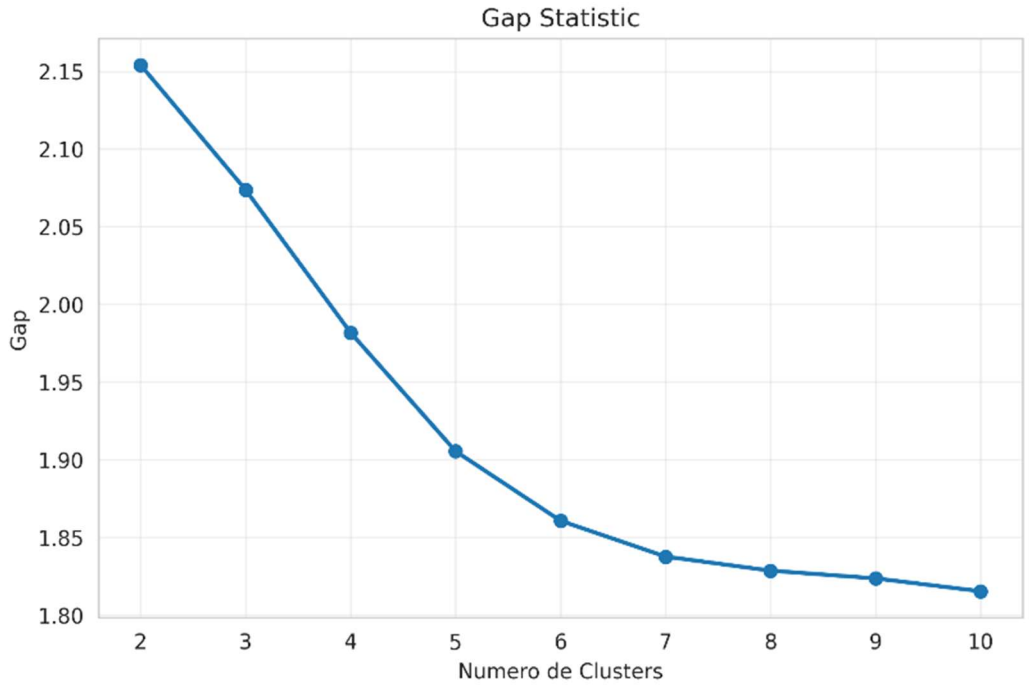
La cantidad de clústeres no se definió utilizando un único indicador. En primer lugar, se evaluaron soluciones comprendidas entre dos y diez grupos para observar cómo cambiaban la separación y la cohesión interna de las empresas. Esta comparación permitió identificar que cada criterio estadístico favorecía una estructura diferente.

Como se observa en la Figura 6, el valor más alto del *Gap Statistic* se presentó cuando $\boldsymbol{K = 2}$. Este resultado sugería una división general de la muestra en dos grupos; sin embargo, dicha solución resumía en exceso las diferencias financieras encontradas entre las empresas. Por esta razón, el resultado no se utilizó de forma aislada.

La decisión final también consideró los índices de *Silhouette* y Davies–Bouldin, así como la posibilidad de interpretar económicamente los conglomerados obtenidos. A partir de esta revisión conjunta se analizaron soluciones con un mayor número de grupos antes de seleccionar la estructura definitiva.

**Figura 6**

*Gap Statistic para la determinación del número óptimo de clústeres*




*Nota*. Fuente: Elaboración propia.

### *4.5.2.* Resultados de validación

Los índices de validación mostraron resultados parcialmente diferentes. Tal como se muestra en la Tabla 12 y en la Figura 7 el índice de Silhouette alcanzó sus mejores valores entre K = 6 y K = 7, mientras que Davies–Bouldin presentó mínimos alrededor de K = 6. No obstante, desde una perspectiva práctica e interpretativa, la solución de tres conglomerados (K = 3) ofreció un equilibrio adecuado entre separación estadística y utilidad analítica.

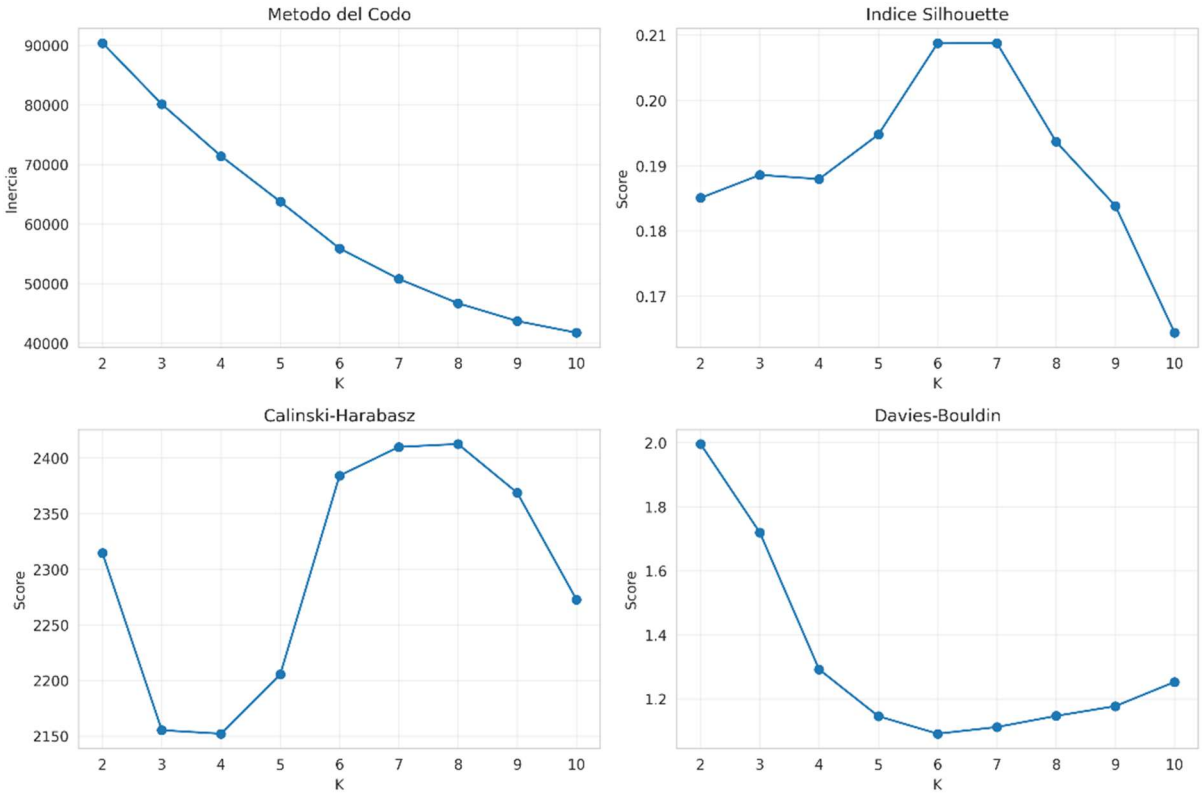
**Tabla 12**

*Resultados de validación del número de clústeres*

| Número de clústeres | Gap Statistic | Silhouette | Davies–Bouldin |
|---|---|---|---|
| 2 | Máximo | Moderado | Moderado |
| 3 | Alto | Adecuado | Adecuado |
| 4 | Intermedio | Adecuado | Adecuado |
| 5 | Intermedio | Alto | Bajo |
| 6 | Intermedio | Alto | Mínimo |
| 7 | Intermedio | Alto | Bajo |

**Figura 7**

*Validación del número de clústeres mediante índices internos.*




Nota. Fuente: Elaboración propia.

### *4.5.3.* Caracterización de los perfiles empresariales

La solución final descrita en la Tabla 13 permitió identificar tres perfiles empresariales diferenciados.

**Tabla 13**

*Perfiles financieros de los clústeres identificados*

| Clúster | Características predominantes |
|---|---|
| Clúster 1 | Empresas con desempeño financiero estable, menor tamaño relativo y estructuras de gasto moderadas. |
| Clúster 2 | Empresas con mayor tamaño económico, elevados niveles de activos y mayores volúmenes operativos. |
| Clúster 3 | Empresas con mayor vulnerabilidad financiera, menor eficiencia operativa y estructuras más frágiles. |

Con fines comparativos, la Tabla 14 muestra un resumen consolidado de las principales características de cada conglomerado.

**Tabla 14**

*Resumen comparativo de los conglomerados finales*

| Característica | Clúster 1 | Clúster 2 | Clúster 3 |
|---|---|---|---|
| Tamaño económico | Medio | Alto | Bajo |
| Eficiencia operativa | Alta | Moderada | Baja |
| Rentabilidad | Positiva | Positiva | Vulnerable |
| Nivel de gastos | Moderado | Alto | Variable |
| Riesgo financiero | Bajo | Moderado | Elevado |

### *4.5.4.* Representación mediante PCA

Con el propósito de facilitar la interpretación visual de los conglomerados obtenidos, se aplicó un Análisis de Componentes Principales (PCA). Proyectar un conjunto de datos que originalmente posee múltiples dimensiones (activos, ingresos, gastos, liquidez, edad, etc.) en un plano bidimensional (2D) es fundamental para observar si los grupos identificados tienen una separación real o si son producto del azar.

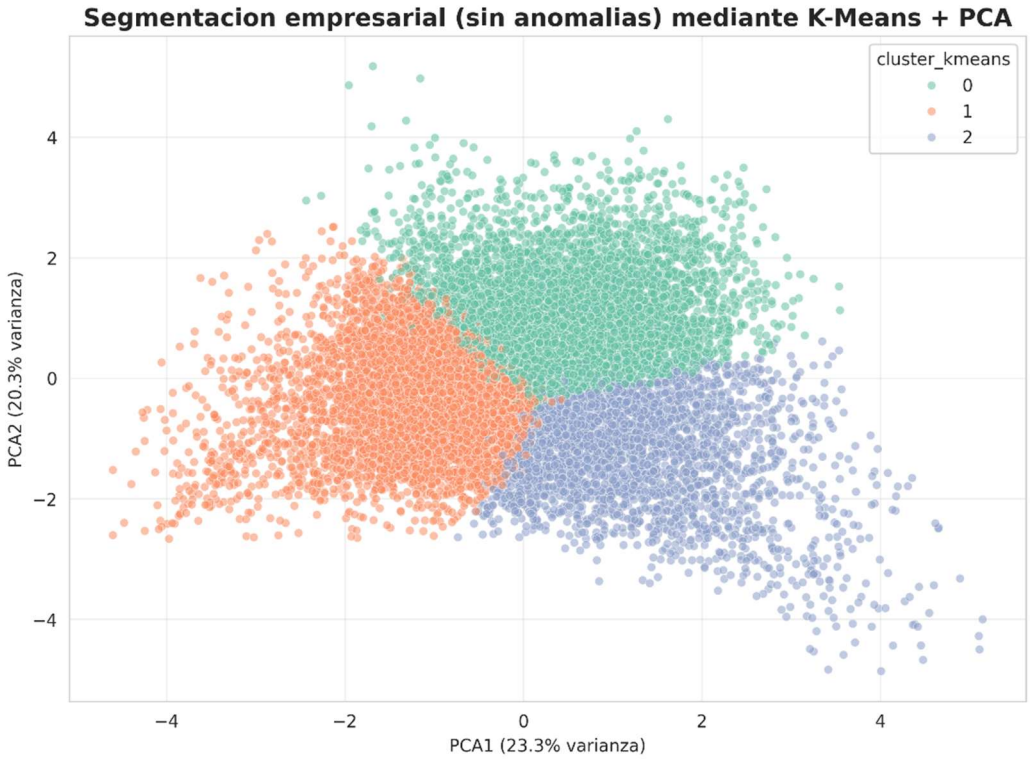
Los dos primeros componentes principales explicaron conjuntamente aproximadamente 43,6 % de la variabilidad total, correspondiendo 23,3 % al primer componente y 20,3 % al segundo componente.

En el contexto del análisis de bases de datos financieras, lograr resumir casi la mitad de la información del mercado nacional en solo dos ejes confirma que los indicadores financieros seleccionados estructuran de forma sólida el perfil de las compañías.

La representación gráfica que se muestra en la Figura 8 evidenció una separación razonable entre los tres grupos identificados, confirmando que las pequeñas empresas ecuatorianas no constituyen un segmento homogéneo desde el punto de vista financiero.

**Figura 8**

*Visualización de los conglomerados mediante PCA y K-Means.*




*Nota*. Fuente: Elaboración propia.

## 4.6. Interfaz gráfica para la predicción del desempeño financiero

Como complemento al pipeline analítico, se implementó una interfaz gráfica en Gradio con el propósito de facilitar el uso de los modelos predictivos por parte de usuarios sin experiencia directa en programación. La aplicación transforma los resultados del notebook en un prototipo interactivo capaz de estimar, a partir de información financiera del año t, la probabilidad de que una pequeña empresa ecuatoriana presente rentabilidad en el período t+1.

### *4.6.1.* Diseño y funcionamiento de la aplicación

Como parte final del desarrollo se consideró conveniente incorporar una interfaz gráfica que permitiera utilizar los modelos predictivos sin tener que interactuar directamente con las celdas de código. Aunque el procesamiento, el entrenamiento y la evaluación de los algoritmos se realizaron en el notebook de Google Colab, la interfaz proporciona una forma más sencilla de ingresar la información de una empresa y consultar el resultado generado por el modelo. Para su implementación se utilizó Gradio, una biblioteca de Python que permite construir aplicaciones interactivas a partir de funciones previamente definidas.

Al iniciar la aplicación, el usuario puede seleccionar uno de los tres modelos supervisados entrenados con la distribución original de los datos: regresión logística, Random Forest o XGBoost. Esta posibilidad de selección resulta útil porque permite observar si la estimación cambia según el algoritmo utilizado. Los tres modelos parten del mismo conjunto de variables

predictoras, pero procesan las relaciones entre ellas de manera diferente. La regresión logística representa un enfoque lineal e interpretable, mientras que Random Forest y XGBoost permiten modelar relaciones más complejas y no lineales.

Después de seleccionar el modelo, el usuario debe ingresar nueve datos relacionados con la situación financiera y operativa de la empresa: ingresos por ventas, activos totales, patrimonio, número de empleados, gastos financieros, gastos administrativos y de ventas, costos de ventas o producción, año fiscal y año de constitución. Estos valores corresponden a información que normalmente puede obtenerse de los estados financieros y de los registros generales de la compañía.

El formulario incluye límites para determinadas variables. Los ingresos por ventas deben encontrarse entre USD 100.001 y USD 1.000.000; los activos totales, entre USD 100.001 y USD 750.000; y el número de empleados, entre 1 y 49. Estos límites no fueron establecidos de manera arbitraria, sino que responden a los criterios utilizados previamente para delimitar el segmento de pequeñas empresas dentro de la base analítica. De esta manera, se procura que los datos ingresados en la interfaz se mantengan dentro de un dominio semejante al empleado durante el entrenamiento de los modelos.

Una vez registrada la información, la aplicación calcula automáticamente otras variables necesarias para realizar la predicción. El pasivo total se obtiene mediante la diferencia entre los activos y el patrimonio. El total de gastos se calcula a partir de la suma de los gastos financieros, los gastos administrativos y de ventas, y los costos de ventas o producción. Posteriormente, la utilidad del ejercicio se estima como la diferencia entre los ingresos por ventas y el total de gastos. La antigüedad empresarial se obtiene restando el año de constitución al año fiscal seleccionado.

La interfaz también construye los indicadores financieros derivados que forman parte del modelo. Entre ellos se encuentran el ratio de endeudamiento, el margen operativo, la rotación de activos y la carga financiera. Asimismo, se generan las transformaciones logarítmicas de los ingresos, los activos y el número de empleados. Estas transformaciones fueron utilizadas durante la preparación de los datos para reducir el efecto de las distribuciones asimétricas y facilitar el procesamiento de variables con escalas diferentes.

Antes de efectuar la predicción, los valores calculados se organizan de acuerdo con el orden establecido en la lista FEATURES del notebook. Este paso es importante porque los modelos esperan recibir las variables con la misma estructura utilizada durante su entrenamiento. En otras palabras, la interfaz no constituye un procedimiento independiente, sino una capa de interacción construida sobre el pipeline analítico desarrollado previamente.

### *4.6.2.* Presentación e interpretación de los resultados

Luego de procesar los datos ingresados, el modelo seleccionado estima la probabilidad de que la empresa pertenezca a la clase rentable en el período siguiente. A partir de esta probabilidad también se genera una clasificación binaria: empresa rentable o empresa con riesgo financiero. El resultado se presenta en un formato resumido para que pueda ser interpretado sin necesidad de revisar directamente los valores internos producidos por el algoritmo.

La primera parte de la respuesta muestra el estado financiero estimado y señala el modelo utilizado. También presenta la probabilidad estimada de rentabilidad futura y su valor complementario como probabilidad de riesgo financiero. Es importante señalar que estos resultados corresponden a una estimación estadística construida a partir de los patrones identificados en los datos históricos. Por tanto, no deben interpretarse como una certeza sobre el comportamiento futuro de la empresa.

Para facilitar la lectura del resultado, la aplicación incorpora un sistema de semáforo. El color verde se utiliza cuando la probabilidad estimada de rentabilidad es igual o superior a 0,80. Este resultado representa el nivel más favorable dentro de la escala definida para la interfaz. El color amarillo corresponde a probabilidades comprendidas entre 0,60 y 0,7999, mientras que el color rojo identifica probabilidades inferiores a 0,60. El semáforo busca ofrecer una referencia visual inmediata, pero no reemplaza la revisión de los indicadores financieros ni el análisis detallado de la empresa.

Además del resumen principal, la aplicación presenta una tabla con los valores calculados durante el procesamiento. En ella se muestran el pasivo total, el total de gastos, la utilidad estimada del ejercicio, el ratio de endeudamiento, el margen operativo, la rotación de activos, la carga financiera y la antigüedad de la empresa. La inclusión de esta tabla permite comprobar cómo fueron transformados los datos originales antes de ser enviados al modelo.

Esta presentación también aporta cierto nivel de transparencia al funcionamiento de la herramienta. El usuario no recibe únicamente una clasificación final, sino que puede revisar los indicadores derivados que intervinieron en la predicción. Esto resulta especialmente relevante en el ámbito financiero, donde una recomendación o una señal de riesgo debe estar acompañada por información que permita comprender el contexto en el que fue generada

**Figura 9**

*Componentes y flujo de la interfaz gráfica de predicción financiera.*


*Nota*. Fuente: Elaboración propia a partir de la aplicación implementada en Gradio.

La Figura 9 resume el funcionamiento general de la aplicación. El proceso comienza con el ingreso de la información financiera y continúa con el cálculo automático de las variables derivadas. Posteriormente, los datos son procesados por el modelo seleccionado y, finalmente, se presentan la clasificación, las probabilidades estimadas, el semáforo y la tabla de indicadores. Este flujo muestra la relación directa entre la interfaz y las etapas de procesamiento implementadas en el notebook.

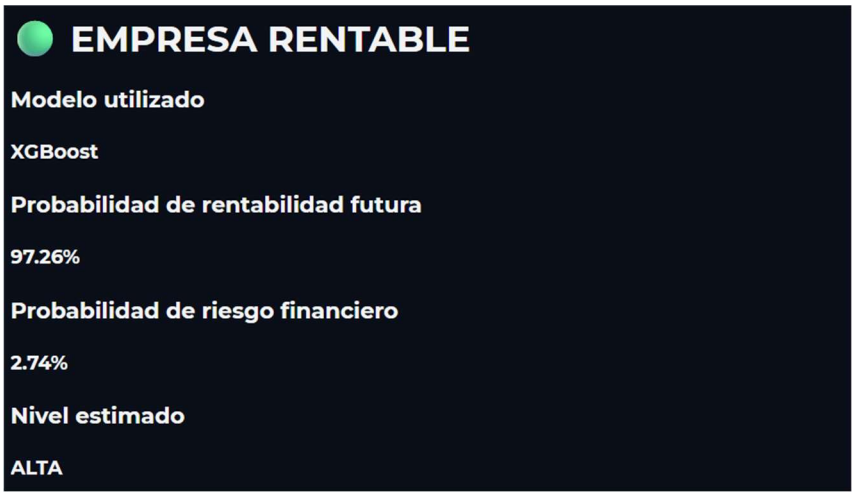
**Figura 10**

*Interfaz gráfica con gradio.*


*Nota*. Fuente: Elaboración propia a partir de la aplicación implementada en Gradio.

**Figura 11**

*Resultados del modelo utilizado a través de un ejemplo de prueba.*




*Nota*. Fuente: Elaboración propia a partir de la aplicación implementada en Gradio.

**Figura 12**

*Resultados de indicadores financieros a través de un ejemplo de prueba.*


*Nota*. Fuente: Elaboración propia a partir de la aplicación implementada en Gradio.

### *4.6.3.* Alcance y condiciones de uso

La interfaz desarrollada constituye un prototipo de apoyo analítico. Para que funcione correctamente, primero deben ejecutarse las etapas de carga, depuración, transformación, entrenamiento y evaluación incluidas en el notebook. Esto se debe a que la aplicación utiliza los modelos almacenados en el objeto modelos_originales, así como la lista FEATURES, generada durante la preparación de la base de modelado. Si estas etapas no han sido ejecutadas previamente, la interfaz no dispone de los objetos necesarios para realizar la predicción.

La aplicación se publica mediante la opción share=True de Gradio. Esta configuración genera un enlace público temporal que permite abrir la herramienta desde un navegador. Sin embargo, dicho enlace permanece disponible únicamente mientras la sesión de Google Colab continúa activa y tiene una duración limitada. Por esta razón, la solución debe entenderse como una demostración funcional y no como una aplicación desplegada permanentemente en un servidor de producción.

Otro aspecto que debe considerarse es que la interfaz utiliza los modelos entrenados con la distribución original de los datos. No se incorporaron en ella los modelos reentrenados

mediante SMOTE+Tomek. Esta decisión permite que el usuario compare las estimaciones generadas por regresión logística, Random Forest y XGBoost bajo el escenario original analizado en el estudio. En consecuencia, el resultado obtenido en la interfaz dependerá tanto de los datos ingresados como del algoritmo seleccionado.

Las estimaciones también deben interpretarse dentro de los límites de la población estudiada. Los modelos fueron desarrollados a partir de información correspondiente a pequeñas empresas ecuatorianas y de variables financieras comprendidas dentro de rangos específicos. Si se ingresan datos alejados de esos valores o pertenecientes a empresas con características muy diferentes, la capacidad de generalización del modelo puede reducirse. Por este motivo, la propia interfaz restringe algunas entradas de acuerdo con los criterios aplicados en la construcción de la base.

Finalmente, el resultado no constituye una certificación de solvencia ni reemplaza el análisis efectuado por especialistas financieros. Tampoco debería utilizarse de manera aislada para aprobar créditos, rechazar solicitudes o tomar decisiones de inversión. Su función principal consiste en mostrar cómo los resultados obtenidos durante la investigación pueden trasladarse hacia una herramienta interactiva y comprensible para el usuario.

En este sentido, la interfaz representa una aplicación práctica del trabajo realizado en el notebook. Además de facilitar la consulta de los modelos, permite demostrar que el pipeline desarrollado puede utilizarse como punto de partida para futuros sistemas de evaluación financiera. Una etapa posterior podría incluir la validación con nuevos períodos, la incorporación de mecanismos adicionales de interpretabilidad y el despliegue de la aplicación en una infraestructura permanente.

# CAPÍTULO 5

# 5. CONCLUSIONES Y RECOMENDACIONES

## 5.1. Conclusiones

El objetivo de este trabajo fue desarrollar un modelo de predicción mediante aprendizaje automático para medir el desempeño financiero futuro de las pequeñas empresas en Ecuador, cuyos resultados demuestran que la información financiera y contable reportada obligatoriamente a la Superintendencia de Compañías, Valores y Seguros (SCVS) poseen información empírica suficiente para anticipar escenarios económicos. Con un poder discriminatorio consistente, el flujo analítico demostró la capacidad de pronosticar la probabilidad de un desempeño favorable en el período siguiente, aportando una herramienta preventiva frente al riesgo de insolvencia.

Asimismo, los datos evidenciaron que un mayor nivel de complejidad computacional no garantiza un rendimiento superior. Si bien el modelo *XGBoost* lideró el rendimiento con la distribución original, el modelo de Regresión Logística alcanzó la mayor efectividad tras aplicar el balanceo con SMOTE+Tomek, lo que confirma que el preprocesamiento de los datos influye de manera determinante en los resultados.

En cuanto a los objetivos específicos, el trabajo consolidó una base de datos financiera robusta y apta para procesamiento avanzado. Paralelamente, la segmentación no supervisada (K-Means y DBSCAN) identificó perfiles empresariales diferenciados, comprobando que las PYMES no responden a un esquema único de riesgo. Al evaluar los modelos supervisados, la investigación reveló un hallazgo metodológico clave: los ensambles complejos (Random Forest y XGBoost) no superaron de forma radical al modelo tradicional de Regresión Logística. Esto indica que la relación entre las variables financieras y el desempeño futuro mantiene una estructura de complejidad moderada, haciendo innecesaria una arquitectura sobredimensionada para lograr clasificaciones precisas.

Por otro lado, el uso de la técnica de remuestreo SMOTE+Tomek generó efectos heterogéneos. Mientras favoreció ligeramente a la Regresión Logística, castigó el rendimiento de los modelos de árboles. Esto demuestra que el balanceo sintético en datos financieros no es una solución universal; la creación de observaciones artificiales puede introducir ruido estadístico que altera el aprendizaje de los algoritmos basados en particiones.

Finalmente, las herramientas de interpretabilidad (*Feature Importance*) confirmaron que el indicador de margen operativo es el determinante absoluto del éxito financiero futuro. Controlar los costos operativos y transformar las ventas en utilidad superó la relevancia de factores tradicionales como el patrimonio, el volumen de activos o la antigüedad. Este hallazgo establece una base sólida para encaminar futuras investigaciones y marcos regulatorios hacia el monitoreo continuo de la eficiencia operativa, facilitando la detección temprana de vulnerabilidades en el tejido empresarial ecuatoriano.

## 5.2. Recomendaciones

A partir de las conclusiones de este trabajo, se proponen las siguientes recomendaciones a fin de orientar futuras investigaciones y facilitar la aplicación práctica de los resultados:

Primero, se sugiere que instituciones financieras como los bancos, cooperativas y organismos de regulación y control utilicen herramientas de análisis predictivo basadas en datos contables oficiales. En la actualidad, la evaluación de crédito en el país se apoya en datos históricos que muestran una realidad de las pequeñas empresas. Al sumar algoritmos de aprendizaje automático con los reportes de la SCVS, estas instituciones podrían complementar sus métodos tradicionales para detectar a tiempo empresas en riesgo y prevenir impagos.

Segundo, se recomienda que futuros estudios amplíen el tiempo de análisis con ciclos económicos más largos. El periodo evaluado sirvió muy bien para pronósticos a corto plazo, pero los mercados locales sufren cambios y crisis externas con frecuencia. Analizar datos de una década entera servirá para comprobar si los modelos (Regresión Logística, *Random Forest y XGBoost*) se mantienen estables frente a crisis o épocas de crecimiento, y si cambian los factores que determinan el éxito financiero con el paso de los años.

Tercero, se sugiere la inclusión de variables macroeconómicas como la inflación, el crecimiento del PIB, la tasa de riesgo país, prácticas de administración, atributos relacionados con el sector, la innovación y uso de tecnología. Aunque la contabilidad es la base de la evaluación, el éxito de un negocio pequeño también depende de su entorno. Combinar estos datos financieros con métricas sobre digitalización o gestión ayudará a que los algoritmos tengan una visión más completa del riesgo empresarial y mejoren sus aciertos.

Cuarto, se propone usar herramientas avanzadas de explicabilidad, como SHAP o LIME. Dado que los modelos de ensamble funcionan con muchos árboles de decisión y su lógica interna es compleja, la utilización y aplicación de estas librerías ayudarán a entender con exactitud las variables que llevaron a una empresa a tener problemas financieros. Esto es fundamental para lograr que los algoritmos sean transparentes y puedan ser aceptados sin problemas en el sector financiero.

Además, se recomienda implementar mecanismos de seguimiento que integren la evolución de indicadores críticos, particularmente el margen operativo. Como quedó demostrado empíricamente en la extracción de la importancia de variables, la eficiencia operativa es el escudo principal contra el fracaso financiero. Utilizar este indicador central como base para sistemas de alerta temprana orientados a la gestión empresarial y la supervisión financiera permitirá a las pequeñas empresas corregir fugas de liquidez y ajustar sus estructuras de costos antes de comprometer la viabilidad de sus negocios en el ejercicio siguiente.

Finalmente, se recomienda promover el uso de metodologías reproducibles basadas en *notebooks* computacionales, documentación exhaustiva y buenas prácticas de ciencia de datos. El desarrollo de proyectos analíticos en el Ecuador debe ir más allá de la presentación final de resultados; requiere que cada etapa de preprocesamiento, balanceo (*SMOTE+Tomek*) y validación cruzada quede debidamente registrada y auditable. La reproducibilidad fortalece la transparencia del proceso investigativo, facilita la validación independiente de los resultados y contribuye a consolidar una cultura científica orientada a

la generación de evidencia robusta y aplicable al contexto empresarial ecuatoriano.

## Bibliografía

Abdel-Aziz, R. I., & Alrabba, H. M. (2023). The Impact of Corporate Governance on Agency Costs in Jordanian Service Companies. *Information Sciences Letters*, *12*(1), 97–111. https://doi.org/10.18576/isl/120107

Abou Elseoud, M. S., Yassin, M., & Ali, M. A. M. (2020). Using a panel data approach to determining the key factors of Islamic banks' profitability in Bahrain. *Cogent Business and Management*, *7*(1). https://doi.org/10.1080/23311975.2020.1831754

Acquah, I. S. K., Quaicoe, J., & Arhin, M. (2022). How to invest in total quality management practices for enhanced operational performance: findings from PLS-SEM and fsQCA. *The TQM Journal*, *35*(7), 1830–1859. https://doi.org/10.1108/TQM-05-2022-0161

Adekunle, B. (2011). Determinants of microenterprise performance in nigeria. *International Small Business Journal*, *29*(4), 360–373. https://doi.org/10.1177/0266242610369751

Ahmeti, A., & Balaj, D. (2023). Influence of Working Capital Management on the SME's Profitability - Evidence from Kosovo. *Quality - Access to Success*, *24*(192), 154–162. https://doi.org/10.47750/QAS/24.192.18

Akinlo, A. E. (2012). Firm size-profitability nexus: Evidence from panel data for Nigeria. *Economic Research*, *25*(3), 706–721. https://doi.org/10.1080/1331677X.2012.11517530

Alarussi, A. S., & Alhaderi, S. M. (2018). Factors affecting profitability in Malaysia. *Journal of Economic Studies*, *45*(3), 442–458. https://doi.org/10.1108/JES-05-2017-0124

Almaqtari, F. A., Al-Homaidi, E. A., Tabash, M. I., & Farhan, N. H. (2019). The determinants of profitability of Indian commercial banks: A panel data approach. *International Journal of Finance and Economics*, *24*(1), 168–185. https://doi.org/10.1002/ijfe.1655

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(1), 193. https://doi.org/10.2307/2325319

Aman, E. E. (2019). Determinants of Financial Distress in Ethiopia Banking Sector. *International Journal of Scientific and Research Publications (IJSRP)*, *9*(5), p8914. https://doi.org/10.29322/ijsrp.9.05.2019.p8914

Balasubramanian, S. A., Radhakrishna, G. S., Sridevi, P., & Natarajan, T. (2019). Modeling corporate financial distress using financial and non-financial variables: The case of Indian listed companies. *International Journal of Law and Management*, *61*(3–4), 457–484. https://doi.org/10.1108/IJLMA-04-2018-0078

Baños-Caballero, S., García-Teruel, P. J., & Martínez-Solano, P. (2014). Working capital management, corporate performance, and financial constraints. *Journal of Business Research*, *67*(3), 332–338. https://doi.org/10.1016/j.jbusres.2013.01.016

Basali, M. (2025). *Impact of Financial Performance and Corporate Governance on ESG Disclosure : Evidence from Saudi Arabia*. 1–19.

Bayaraa, B. (2017). Financial Performance Determinants of Organizations: The Case of Mongolian Companies. *Journal of Competitiveness*, *9*(3), 22–33. https://doi.org/10.7441/joc.2017.03.02

Besley, S., & Brigham, E. F. (2001). *Fundamentos de Administración Financiera* (A. González Maya (ed.); Décimosegu). McGraw-Hill Interamericana Editores S.A.

Bozkurt, İ., & Kaya, M. V. (2022). Foremost features affecting financial distress and Bankruptcy in the acute stage of COVID-19 crisis. *Applied Economics Letters*. https://doi.org/10.1080/13504851.2022.2036681

Brealey, R. A., Myers, S. C., & Allen, F. (2010). *Principios de Finanzas Corporativas* (Novena). McGraw-Hill Interamericana Editores S.A. https://www.u-cursos.cl/usuario/b8c892c6139f1d5b9af125a5c6dff4a6/mi_blog/r/Principios_de_Finanzas_Corporativas_9Ed__Myers.pdf

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Burja, C. (2011). Factors Influencing The Companies‘ Profitability. *Annales Universitatis Apulensis Series Oeconomica*, *2*(13), 215–224. https://doi.org/10.29302/oeconomica.2011.13.2.3

Campillo, J. P., Vargas, J. M., & Ibáñez, P. C. (2018). Analysis of the algorithm Gradient Boosting Machine (GBM) in business failure prediction [Análisis de la utilidad del algoritmo Gradient Boosting Machine (GBM) en la predicción del fracaso empresarial].

*Revista Espanola de Financiacion y Contabilidad*, *47*(4), 507–532. https://doi.org/10.1080/02102412.2018.1442039

CEPAL. (2009). *Manual de la Micro, Pequeña y Mediana Empresa*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. http://arxiv.org/abs/2003.09788

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.

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications Inc. https://doi.org/10.1016/j.math.2010.09.003

Cultrera, L., & Brédart, X. (2016). Bankruptcy prediction: The case of Belgian SMEs. *Review of Accounting and Finance*, *15*(1), 101–119. https://doi.org/10.1108/RAF-06-2014-0059

Dia, M., Takouda, P. M., & Golmohammadi, A. (2022). Assessing the performance of Canadian credit unions using a three-stage network bootstrap DEA. *Annals of Operations Research*, *311*(2), 641–673. https://doi.org/10.1007/s10479-020-03612-w

Dieperink, H., Adriaanse, J., & Dechesne, M. (2025). Predicting viability of small businesses on the edge of failure. *Journal of Small Business Management*, *63*(5), 2422–2454. https://doi.org/10.1080/00472778.2024.2435506

Durica, M., Valaskova, K., & Janoskova, K. (2019). Logit business failure prediction in V4 countries. *Engineering Management in Production and Services*, *11*(4), 54–64. https://doi.org/10.2478/emj-2019-0033

Durrah, O., Rahman, A. A. A., Jamil, S. A., & Ghafeer, N. A. (2016). Exploring the relationship between liquidity ratios and indicators of financial performance: An analytical study on food industrial companies listed in Amman Bursa. *International Journal of Economics and Financial Issues*, *6*(2), 435–441.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for

Discovering Clusters in Large Spatial Databases with Noise. *Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, *2*, 226–231. https://doi.org/10.1016/B978-0-444-64165-6.03005-6

García Pérez de Lema, D., & Sánchez Ballesta, J. (2003). Influencia del tamaño y la antigüedad de la empresa sobre la rentabilidad: un estudio empírico. In *Revista de Contabilidad : Spanish Accounting Review* (Vol. 6, Issue 12, pp. 169–206).

Garg, V., Tewari, P., & Srivastav, S. (2018). Liquidity and profitability analysis of selected automobile companies. *International Journal of Supply Chain Management*, *7*(4), 101–110.

Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting*, *29*(6), 536–555. https://doi.org/10.1002/for.1153

Gołaś, Z. (2020). The effect of inventory management on profitability: evidence from the Polish food industry: Case study. *Agricultural Economics (Czech Republic)*, *5*, 234–242.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hernández Sampieri Roberto, Fernández Collado Carlos, & Baptista Lucio Pilar. (2014). *Metodología de la investigación* (6th ed.). McGraw Hill España.

Ho, C. S. F., & Mohd-Raff, N. E. N. (2019). External and internal determinants of performances of Shariah and non-Shariah compliant firms. *International Journal of Islamic and Middle Eastern Finance and Management*, *12*(2), 236–253. https://doi.org/10.1108/IMEFM-08-2017-0202

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression: Third Edition. In *Applied Logistic Regression: Third Edition*. https://doi.org/10.1002/9781118548387

Hossain, T. (2021). Determinants of profitability: A study on manufacturing companies listed on the dhaka stock exchange. *Asian Economic and Financial Review*, *10*(12), 1496–1508. https://doi.org/10.18488/JOURNAL.AEFR.2020.1012.1496.1508

INEC. (2017a). *Panorama laboral y empresarial del Ecuador*.

INEC. (2017b). Supervivencia Empresarial: Factores asociados al cierre de empresas del sector productivo ecuatoriano en el periodo 2009-2015. In *Instituto Nacional de Estadísticas y Censos*. https://www.ecuadorencifras.gob.ec/documentos/web-inec/boletin/Presentaciones_Seminario_Sec_Lab/Supervivencia_empresarial_factores_asociados_a_la_muerte_de_empresas_en_Ecuador.pdf

Instituto Nacional de Estadística y Censos - INEC. (2020). Directorio de Empresas y Establecimientos 2019. In *Ecuador en Cifras*.

Jamali, A. H. (2012). Management efficiency and profitability in Indian automobile industry : from theory to practice. *Indian Journal of Science and Technology*, *5*(5), 2779–2781.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning, Springer Texts. In *Springer Texts*.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Lee, S. (2014). The relationship between growth and profit: Evidence from firm-level panel data. *Structural Change and Economic Dynamics*, *28*, 1–11. https://doi.org/10.1016/j.strueco.2013.08.002

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings Ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, *1*, 281–297. https://doi.org/10.1007/s11665-016-2173-6

Mahmood, F., Ahmed, Z., Hussain, N., & Ben-zaied, Y. (2025). Working capital financing and firm performance Working capital financing and firm performance : a machine learning approach. In *Review of Quantitative Finance and Accounting* (Vol. 65, Issue 1).

Springer US. https://doi.org/10.1007/s11156-023-01185-w

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable . christophm. github. io/interpretable-ml-book. In *Samek, W., Montavon, G., Lapuschkin, S., Anders, CJ, & Müller, KR (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE* (Vol. 109, Issue 3).

Nimer, M. Al, Warrad, L., & Omari, R. Al. (2015). The Impact of Liquidity on Jordanian Banks Profitability through Return on Assets. *European Journal of Business and Management*, *7*(7), 229–233.

Nindita, K. (2014). Prediction on Financial Distress of Mining Companies Listed in BEI using Financial Variables and Non-Financial Variables. *European Journal of Business and Management*, *6*(34), 226–237.

Nurhayati, Mufidah, A., & Kholidah, A. N. (2017). The Determinants of Financial Distress of Basic Industry and Chemical Companies Listed in Indonesia Stock Exchange. *Review Of Management And Entreprenuership*, *01*(02/2017), 19–26.

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109. https://doi.org/10.2307/2490395

Parra, J. F. (2011). PROBABILITY DETERMINANTS OF NEW ENTERPRISES CLOSURE IN BOGOTÁ. *Rev.Fac.Cienc.Econ.*, *XIX*(1), 27–53.

Parrales Choez, C. G., Zambrano Farías, F. J., & Valls Martínez, M. del C. (2024). Determinants of the profitability of credit unions in Ecuador. *Journal of Infrastructure, Policy and Development*, *8*(8), 1–15.

Pietrzak, M. (2022). Can financial sector distress be detected early? *Borsa Istanbul Review*. https://doi.org/10.1016/j.bir.2022.08.002

Provost & Fawcett. (2013). *Data science-what you need to know about analytic-thinking and decision-making*.

Quispe-Fernández, G. M., & Ayaviri-Nina, D. (2021). Carga y presión tributaria. Un estudio

del efecto en la liquidez, rentabilidad e inversión de los contribuyentes en Ecuador. *Retos*, *11*(22), 247–263.

Rafatnia, A. A., Ramakrishnan, S., Abdullah, D. F. B., Nodeh, F. M., & Farajnezhad, M. (2020). Financial distress prediction across firms. *Journal of Environmental Treatment Techniques*, *8*(2), 646–651.

Ragab, Y. M., & Saleh, M. A. (2021). Non-financial variables related to governance and financial distress prediction in SMEs–evidence from Egypt. *Journal of Applied Accounting Research*, *23*(3), 604–627. https://doi.org/10.1108/JAAR-02-2021-0025

Rahman, S. M. M., & Saif, A. N. M. (2021). Performance evaluation of some listed companies: Insights from Dhaka Stock Exchange in Bangladesh. *Global Business and Economics Review*, *24*(3), 261–278. https://doi.org/10.1504/GBER.2021.114661

Rao, Z.-U.-R., Huzaifa, M., & Safdar, R. (2025). Capital structure and firm performance: evidence from energy sector. *Journal of Social and Economic Development*, *27*(3), 846–862. https://doi.org/10.1007/s40847-025-00442-z

Romero Espinosa, F. (2013). Variables financieras determinantes del fracaso empresarial para la pequeña y mediana empresa en Colombia : análisis bajo modelo Logit. *Pensamiento & Gestion*, *34*, 235–277.

Romero Martínez, M., Carmona Ibáñez, P., & Pozuelo Campillo, J. (2021). La predicción del fracaso empresarial de las cooperativas españolas. Aplicación del Algoritmo Extreme Gradient Boosting. In *CIRIEC-Espana Revista de Economia Publica, Social y Cooperativa* (Vol. 101). https://doi.org/10.7203/CIRIEC-E.101.15572

Ross, S. A., Westerfield, R. W., & Jordan, B. D. (2010). *Fundamentos de Finanzas Corporativas* (M. Á. Toledo Castellano (ed.); Novena). McGraw-Hill Educación.

Sánchez Pacheco, M. E., Bermúdez Fajardo, P. N., Zea Franco, R. D., & Zambrano Farías, F. J. (2022). Liquidez, endeudamiento y rentabilidad de las mipymes en Ecuador: un análisis comparativo. *INNOVA Research Journal*, *7*(3.2), 36–50. https://doi.org/10.33890/innova.v7.n3.2.2022.2209

Sánchez Pacheco, M. E., Valls Martínez, M. del C., & Zambrano Farías, F. J. (2025).

Incidencia del capital circulante en la rentabilidad de las microempresas en Ecuador. *Retos Revista de Ciencias Administrativas y Economía*, *15*(2025), 327–340.

Sehgal, S., Mishra, R. K., Deisting, F., & Vashisht, R. (2021). On the determinants and prediction of corporate financial distress in India. *Managerial Finance*, *47*(10), 1428–1447. https://doi.org/10.1108/MF-06-2020-0332

Shahnia, C., Purnamasari, E. D., Hakim, L., & Endri, E. (2020). Determinant of profitability: Evidence from trading, service and investment companies in Indonesia. *Accounting*, *6*(5), 787–794. https://doi.org/10.5267/j.ac.2020.6.004

Shin, G. H. (2008). The profitability of asset sales as an explanation of asset divestitures. *Pacific Basin Finance Journal*, *16*(5), 555–571. https://doi.org/10.1016/j.pacfin.2007.10.004

Singapurwoko, A., & El-Wahid, M. S. M. (2011). The impact of financial leverage to profitability study of non-financial companies listed in Indonesia stock exchange. *European Journal of Economics, Finance and Administrative Sciences*, *32*, 136–148.

Steinerowska-Streb, I. (2012). The determinants of enterprise profitability during reduced economic activity. *Journal of Business Economics and Management*, *13*(4), 745–762. https://doi.org/10.3846/16111699.2011.645864

Stephen, M., & Elvis, K. (2011). Influence of working capital management on firms profitability: A case of smes in Kenya. In *International Business Management* (Vol. 5, Issue 5, pp. 279–286). https://doi.org/10.3923/ibm.2011.279.286

Takon, S. M., & Atseye, F. A. (2015). Effect of working capital management on firm profitability in selected Nigerian quoted companies. *International Journal of Economics, Commerce and Management*, *III*(10), 414–438.

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

Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man, and*

*Cybernetics*, *SMC-6*(11), 769–772. https://doi.org/10.1109/TSMC.1976.4309452

Valls Martínez, M. del C., Santos-Jaén, J. M., Sánchez Pacheco, M. E., & Zambrano Farías, F. J. (2026). Do Governance Structures Drive Green Building Adoption ? A Machine Learning Approach With Random Forests. *Business Strategy and the Environment*, 1–18. https://doi.org/10.1002/bse.70641

Van Horne, J. C., & Wachowicz, J. M. (2010). *Fundamentos de Administración Financiera* (G. Dominguez Chávez & F. Hernández Carrasco (eds.); Décimoterc). Pearson Educación S.A.

Vašaničová, P., Košíková, M., Jenčová, S., Miškufová, M., & Korečko, J. (2025). Financial Performance-Based Clustering of Spa Enterprises in Slovakia. In *Journal of Risk and Financial Management* (Vol. 18, Issue 9, p. 482). https://doi.org/10.3390/jrfm18090482

Vatavu, S. (2014). The determinants of profitability in companies. *Annals of the University of Petrosani, Economics*, *14*(1), 329–338.

Welsh, D. H. B., Munoz, J. M., Deng, S., & Raven, P. V. (2013). Microenterprise performance and microenterprise zones (MEZOs) in China. *Management Decision*, *51*(1), 25–40. https://doi.org/10.1108/00251741311291292

Wisna, N., Listianingsih, M., Al-Azhary, S. N., Kastaman, & Kotjoprayudi, R. B. (2023). Implementation of K-Means Algorithm Using Phyton Based on Profitability and Solvency Ratio. *Journal of Theoretical and Applied Information Technology*, *101*(19), 6200–6206.

Zainudin, N. S., Ting, C. Y., Khor, K. C., Ng, K. H., Tong, G. K., & Kalid, S. N. (2024). Clustering Defensive Shariah-compliant Stocks Using Financial Performance as the Indicator. *International Journal on Informatics Visualization*, *8*(1), 115–122. https://doi.org/10.62527/joiv.8.1.2269

Zambrano-Farías, F. J., Rivera-Naranjo, C. I., & Sánchez-Pacheco, M. E. (2023). Rentabilidad de las mipymes del sector inmobiliario en Ecuador. *Revista Venezolana de Gerencia*, *28*(103), 1021–1036.

Zambrano-Farías, F. J., Sánchez-Pacheco, M. E., Martinez Mayorga, R. X., & Guarnizo

Crespo, S. F. (2022). Determinantes de la rentabilidad financiera del sector comercio: Un estudio transversal para el sector comercio. *Universidad y Sociedad*, *14*(6), 625–632.

Zambrano-Farías, F. J., Sánchez-Pacheco, M. E., Martínez Mayorga, R. X., & Guarnizo Crespo, S. F. (2022). Determinantes de la rentabilidad financiera del sector comercio: Un estudio transversal. *Universidad y Sociedad*, *14*(6), 625–632.

Zambrano-Farías, F. J., Valls-Martínez, M. del C., & Sanchez-Pacheco, M. E. (2024). Determinantes de la insolvencia financiera en empresas de piedra natural en España e Italia. *Estudios Gerenciales*, *40*(172), 271–282.

Zambrano Farías, F. J., Sánchez Pacheco, M. E., & Correa Soto, S. R. (2021). Análisis de rentabilidad, endeudamiento y liquidez de microempresas en Ecuador. *Retos*, *11*(22), 235–249.

Zambrano Farías, F. J., Sánchez Pacheco, M. E., & Valls Martínez, M. del C. (2021). Factors Explaining the Business Survival of MSMEs in Ecuador. *Studies of Applied Economics*, *39*(8), 1–18. https://doi.org/10.25115/eea.v39i8.4061

Zhao, X., Yeung, K., Huang, Q., & Song, X. (2015). Improving the predictability of business failure of supply chain finance clients by using external big dataset. *Industrial Management and Data Systems*, *115*(9), 1683–1703. https://doi.org/10.1108/IMDS-04-2015-0161