Bases teóricas y metodológicas para el desarrollo de un modelo predictivo que identifica la creación y éxito de nuevas PYMES

Ricardo López Gómez

Asesor

Dra. Nidia Danigza Lugo López

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

El presente trabajo aborda la problemática de la alta tasa de fracaso de Pequeñas y Medianas Empresas (PYMES) y propone un conjunto de lineamientos teóricos y metodológicos para el desarrollo de un modelo predictivo de Machine Learning (ML) que determine la probabilidad de creación y éxito inicial de nuevas empresas. Lo anterior, se lleva a cabo mediante la metodología de Revisión Sistemática de la Literatura (RSL), y se clasifican las variables geográficas y de consumo poblacional, estableciendo su jerarquía de significancia para la estructura de la futura base de datos. Adicionalmente, se sintetizan los modelos de Machine Learning más empleados en estudios sobre creación y éxito de PYMES en Colombia. Como resultado, se diseña un marco detallado de lineamientos en 6 fases que garantiza la consecución futura del modelo predictivo, el cual está orientando a emprendedores, inversionistas y demás interesados que les permita identificar oportunidades de mercado y el diseño de estrategias basadas en evidencia geoespacial y de consumo.

***Palabras clave*:** PYMES, Machine Learning, mercado, geoespacial, emprendimiento.

Abstract

This study addresses the problem of the high failure rate among Small and Medium-sized Enterprises (SMEs) and proposes a set of theoretical and methodological guidelines for developing a Machine Learning predictive model that determines the probability of creation and initial success of new enterprises. This is accomplished through a systematic literature review methodology, whereby geographic and population consumption variables are classified, establishing their significance hierarchy for the structure of the future database. Additionally, the most commonly employed Machine Learning models in studies on SME creation and success in Colombia are synthesized. As a result, a detailed framework of guidelines is designed to ensure the future development of the predictive model, which is aimed at entrepreneurs, investors, and other stakeholders, enabling them to identify market opportunities and design evidence-based strategies using geospatial and consumption analytics.

Keywords**:** SMEs, Machine Learning, market, geospatial, entrepreneurship

Tabla de Contenido

Lista de Tablas

Introducción

En Colombia, las Pequeñas y Medianas Empresas (PYMES) constituyen el 99,3% del tejido empresarial y generan aproximadamente el 80% del empleo formal (Confecámaras, 2023; DANE, 2023). No obstante, la tasa de supervivencia a cinco años apenas alcanza el 33,5%, situándose 6,5 puntos porcentuales por debajo del promedio de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) (Confecámaras, 2023). Esta mortalidad empresarial, tradicionalmente atribuida a factores financieros y de gestión, revela una dimensión subestimada, la desarticulación entre la oferta empresarial y las necesidades de consumo territoriales específicas. La heterogeneidad geográfica colombiana, donde el 62,3% de nuevas empresas se concentra en tres departamentos (Confecámaras, 2023); además, la Comisión Económica para América Latina y el Caribe (CEPAL) indica que la infrautilización de tecnologías predictivas con inversión en Inteligencia Artificial (IA) proyectada en $200 millones para 2025 evidencian una brecha metodológica crítica (CEPAL, 2024).

De igual manera, el Machine Learning (ML) ha demostrado reducir errores de clasificación hasta 30% comparado con métodos tradicionales (Alfaro et al., 2008; Romero et al., 2024), la literatura nacional carece de marcos metodológicos integrados que sistematicen variables geoespaciales y de consumo para predecir el éxito de PYMES. Los estudios existentes analizan factores de manera aislada, o aplican modelos genéricos sin adaptación al contexto colombiano, donde el 92% de PYMES son familiares y enfrentan una carga tributaria del 33,4% (ANIF, 2024). Esta investigación aborda dicha brecha mediante una Revisión Sistemática de la Literatura (RSL) que sintetiza evidencia internacional y la contextualiza con datos geoespaciales del Sistema Nacional de Información Empresarial (DANE, 2024).

El objetivo general consiste en proponer lineamientos teóricos y metodológicos para el desarrollo de un modelo predictivo de Machine Learning que determine la probabilidad de creación y éxito inicial de nuevas PYMES, integrando variables geográficas y de consumo poblacional. Esto implica, clasificar variables predictoras según significancia estadística y práctica; sintetizar algoritmos de Machine Learning validados para contextos latinoamericanos; y, diseñar los lineamientos que permitan replicar el desarrollo futuro de modelos predictivos. Por ende, el presente estudio se circunscribe a la fase de diseño conceptual y la implementación computacional constituirá un trabajo futuro.

Este trabajo se organiza en seis secciones, el marco teórico fundamenta conceptualmente el Machine Learning, el emprendimiento y las variables predictoras; la metodología detalla la Revisión Sistemática de Literatura (RSL); los resultados presentan la taxonomía de variables, síntesis de modelos y lineamientos; así mismo, la discusión contrasta hallazgos con la realidad colombiana; y, finalmente, las conclusiones sintetizan aportes y limitaciones.

Planteamiento del Problema

En el emprendimiento, la creación de nuevas Pequeñas y Medianas Empresas (PYMES) son reconocidas como pilares fundamentales para el dinamismo económico, la generación de empleo y la innovación en diversos contextos geográficos. Sin embargo, las nuevas PYMES frecuentemente enfrentan desafíos significativos que resultan en altas tasas de fracaso durante sus etapas iniciales (OECD, 2022). Por ello, algunas investigaciones recientes muestran que las causas primordiales de este fenómeno radican en la poca comprensión de las dinámicas del mercado y la falta de información entre la oferta de la nueva empresa y las necesidades particulares de los consumidores en un entorno espacial dado (Stam, 2021).

De acuerdo con el estudio de supervivencia empresarial realizado por Confecámaras (2023), la tasa de supervivencia de las empresas colombianas a cinco años es tan solo del 33,5%, evidenciando que, de 296.896 unidades productivas creadas en 2017, únicamente 98.696 continuaban operando en 2022. Esta cifra se ubica por debajo del promedio de países de la OCDE, donde la supervivencia supera el 40%, y resulta particularmente crítica en microempresas (33,4%) comparado con empresas medianas (73,7%) (Confecámaras, 2023).

A nivel regional, la Comisión Económica para América Latina y el Caribe CEPAL (2024) advierte que la infrautilización de tecnologías predictivas profundiza las brechas de desarrollo empresarial en la región, donde se proyecta que la inversión en Inteligencia Artificial alcanzará $5,4 mil millones en 2025, pero con una adopción desigual entre países.

Ahora bien, en el panorama colombiano, las PYMES constituyen una porción considerable del universo empresarial (98%), lo cual contribuye al producto interno del país y a la empleabilidad a nivel nacional (de Lema & Segura, 2021). Así mismo, la diversidad regional en términos de desarrollo económico, infraestructura, aspectos culturales, y patrones de consumo

generan retos específicos para el surgimiento y fortalecimiento de nuevas empresas. Por lo tanto, comprender estas dinámicas es sumamente relevante para identificar las necesidades poblacionales e incrementar la probabilidad de éxito de las nuevas PYMES (OECD, 2022).

Por otro lado, el Banco de la República (2024) ha demostrado que variables de consumo poblacional como la variación del crédito al consumo, el Índice de Confianza del Consumidor (ICC) y los indicadores geográficos de desempleo explican más del 70% de las variaciones del consumo en Colombia, subrayando la relevancia de integrar estos factores en modelos predictivos empresariales. Además, estudios de supervivencia empresarial confirman que la localización geográfica es determinante; así empresas ubicadas en zonas estratégicas urbanas y departamentos más desarrollados tienen menor probabilidad de fracaso, así mismo, que la existencia de clústeres productivos aumenta significativamente las tasas de supervivencia (Confecámaras, 2023). No obstante, la mayoría de las investigaciones previas han abordado estos factores de manera aislada, sin integrar variables geográficas y de consumo en modelos predictivos específicos para el contexto colombiano.

Además, la diversidad regional en términos de desarrollo económico, infraestructura, aspectos culturales y patrones de consumo genera retos específicos para el surgimiento de nuevas empresas (Díaz-Martínez et al., 2004). Sin embargo, existe una brecha metodológica crítica, donde el Machine Learning ha demostrado superioridad en la predicción de fracaso empresarial con algoritmos como XGBoost, que reducen errores de clasificación hasta en 30% comparado con métodos tradicionales (Alfaro et al., 2008), la literatura académica carece de lineamientos sistemáticos que integren variables geoespaciales y de consumo poblacional para el contexto colombiano. Romero et al. (2024) señalan que la falta de transparencia en modelos predictivos y

la ausencia de marcos metodológicos adaptados a la realidad latinoamericana limitan la aplicación práctica para emprendedores e inversionistas.

Con base en lo mencionado anteriormente, se plantea la siguiente pregunta de investigación que pretende mitigar este fenómeno.

¿Qué lineamientos teóricos y metodológicos se pueden proponer para el desarrollo de un futuro modelo predictivo de Machine Learning que determine la probabilidad de creación y éxito inicial de nuevas Pequeñas y Medianas Empresas (PYMES), analizando la influencia de variables geográficas y de consumo poblacional?

Justificación

La tasa de mortalidad de las nuevas PYMES representa un desafío significativo para el desarrollo económico y social, tanto a nivel global como en el contexto colombiano. Esta situación genera pérdidas económicas, desaprovechamiento de potencial innovador y frustración de iniciativas emprendedoras (Roman, 2021). Comprender los factores que influyen en el éxito o creación de estas empresas, especialmente en sus etapas iniciales, es imprescindible para diseñar estrategias que busquen la sostenibilidad y el crecimiento.

Además, existe una limitada comprensión de cómo la interacción entre la ubicación geográfica y las necesidades de consumo específicas de la población impacta en el potencial de éxito de nuevas PYMES. Si bien existen estudios que analizan el impacto de factores geográficos o de las características del consumidor, una comprensión integrada de estas dos dimensiones y su capacidad predictiva para la creación de empresas aun requiere mayor exploración (Njeru, 2023).

Así mismo, las PYMES representan el 99% del tejido empresarial colombiano y generan aproximadamente el 80% de los empleos formales, por lo que su fracaso masivo afecta directamente la estabilidad de 3,2 millones de familias dependientes de estos ingresos (DANE, 2023; La Nota Económica, 2024).

Las causas estructurales de esta mortalidad empresarial están intrínsecamente ligadas a una comprensión incompleta de las dinámicas territoriales y de consumo específicas del contexto colombiano. Aunque existen investigaciones que analizan factores geográficos o características del consumidor por separado, la literatura académica nacional carece de un marco integrado que combine ambas dimensiones mediante técnicas predictivas avanzadas (Njeru, 2023). Esta brecha metodológica se acentúa por la heterogeneidad regional de Colombia; el 62,3% de nuevas empresas se concentran en Bogotá, Antioquia y Valle del Cauca, mientras que regiones como la

Guajira, Vaupés y Vichada registran densidades empresariales inferiores a 5 empresas por mil habitantes, evidenciando barreras de acceso a información de mercado calibrada territorialmente (Confecámaras, 2023; MDc, 2025). Asu vez, el Banco de la República (2024) demuestra que variables de consumo poblacional como el crédito al consumo, el Índice de Confianza del Consumidor y desempleo regional explican más del 70% de las variaciones del consumo en Colombia, subrayando que la falta de integración de estos datos en modelos predictivos constituye una oportunidad perdida para anticipar demandas de mercado.

Por ende, el presente trabajo pretende abordar algunos vacíos identificados. Primero, aunque el Machine Learning ha demostrado superioridad en predicción de fracaso empresarial (reduciendo errores de clasificación hasta 30% contra métodos tradicionales), la mayoría de los estudios aplican marcos genéricos sin adaptación a las particularidades del contexto colombiano, donde el 92% de las PYMES son de origen familiar y enfrentan obstáculos únicos como carga tributaria del 33,4% y burocracia que demora 11,3 días para iniciar operaciones (ANIF, 2024; Romero et al., 2024). Segundo, la infraestructura de datos colombiana fortalecida por la Ley 1712 de 2014 y el Sistema Nacional de Información Empresarial del DANE ofrece conjuntos de datos geoespaciales y de consumo subutilizados, mientras que la inversión en Inteligencia Artificial (IA) en el país proyecta alcanzar $200 millones en 2025 con adopción desigual entre regiones (CEPAL, 2024; Infraestructura de Datos, 2024). Tercero, las soluciones existentes se concentran en financiamiento y digitalización básica (fintech como Sempli y Tiendanube cubren solo 31% de PYMES con CRM), pero no integran análisis predictivo territorial y de consumo para identificar oportunidades de nicho antes de la inversión inicial (Bancolombia, 2024; Godaddy, 2024).

Por lo tanto, al proponer lineamientos teóricos y metodológicos que integren variables geográficas y de consumo poblacional mediante Machine Learning representa una contribución científica y práctica focalizada. Desde el ámbito académico, esta investigación sistematiza modelos Machine Learning validados internacionalmente (XGBoost, Random Forest, entre otros) con variables contextuales colombianas, generando un marco reproducible para otros países latinoamericanos con brechas similares. Para el ecosistema empresarial colombiano, estos lineamientos constituirían el primer paso metodológico para desarrollar futuros modelos predictivos que permitan evaluar probabilidades de éxito con granularidad municipal, priorizar sectores con mayor tasa de supervivencia, y diseñar estrategias basadas en mapas de demanda latente. Para formuladores de política pública, este marco orientaría la asignación de recursos a territorios con mayor potencial de retorno, alineando el fomento empresarial con datos basados en evidencia; contribuyendo así a la equidad territorial y consolidación de la clase media emprendedora, objetivos centrales del Plan Nacional de Desarrollo (PND) con vigencia 2022-2026.

Objetivos

Objetivo General

Proponer un conjunto de lineamientos teóricos y metodológicos detallados para el desarrollo de un modelo predictivo de Machine Learning que determine la probabilidad de creación y éxito inicial en Pequeñas y Medianas Empresas (PYMES), analizando la influencia de variables geográficas y de consumo poblacional.

Objetivos Específicos

Clasificar las variables geográficas y de consumo poblacional, estableciendo su jerarquía de significancia que fundamenten la estructura de la base de datos del futuro modelo predictivo, desde la revisión de la literatura.

Sintetizar los modelos de Machine Learning que más se emplean en la literatura en el estudio de las variables que influyen en la creación y éxito de las Pymes existentes en Colombia.

Diseñar el conjunto de lineamientos metodológicos para el desarrollo del modelo predictivo que garantiza su consecución.

Marco Teórico

Fundamentos Teóricos del Machine Learning para la Predicción Empresarial

Un modelo de Machine Learning (ML) es una representación matemática construida a partir de datos, que le permiten a un sistema aprender patrones complejos y realizar predicciones y tomar decisiones sin ser programado explícitamente para cada tarea (Bishop, 2023). A diferencia de los modelos estadísticos tradicionales, los modelos de ML están diseñados para manejar grandes volúmenes de datos, capturar relaciones no lineales y adaptarse a estructuras de alta dimensionalidad, lo que los hace especialmente útiles en contextos empresariales dinámicos y multifactoriales (Goodfellow et al., 2016; Hastie et al., 2021).

En el ámbito de la predicción del éxito de las PYMES, un modelo de ML no solo identifica relaciones entre variables predictoras (geográficas, y de consumo) y el resultado (éxito o fracaso), sino que también es capaz de generalizar estos patrones a nuevos casos no vistos durante el entrenamiento.

La relevancia de emplear modelos de ML en lugar de métodos clásicos radica en su capacidad para:

- Manejar grandes volúmenes de datos (estructurados y no estructurados).
- Detectar interacciones complejas entre variables que un análisis lineal pasaría por alto.
- Ofrecer escalabilidad y adaptabilidad a medida que nuevos datos estén disponibles (Razaghzadeh Bidgoli et al., 2024).

Es fundamental, que estos modelos sean entrenados con datos de calidad, representativos y actualizados, ya que su capacidad predictiva está directamente ligada a la integridad y relevancia de los datos utilizados (Bishop, 2023).

Si bien los modelos más complejos pueden capturar relaciones no lineales, la regresión logística sirve como una línea base fundamental. Su rendimiento se utilizará para evaluar si algoritmos más sofisticados como Random Forest, XGBoost o Redes Neuronales aportan una mejora predictiva significativa que justifique su mayor complejidad y naturaleza (Molitor et al., 2023). Esta aproximación metodológica escalonada asegura un desarrollo robusto y comprensible del modelo predictivo final.

A su vez, la teoría del aprendizaje estadístico proporciona el marco formal para comprender cómo los algoritmos de ML generalizan patrones a partir de datos históricos para realizar predicciones sobre datos no vistos (Hastie et al., 2021). Esta teoría es crucial en el desarrollo de un modelo predictivo confiable, ya que establece conceptos fundamentales como el sesgo-varianza, que determina la capacidad de un modelo al capturar la señal subyacente en los datos sin sobreajustarse al ruido (James et al., 2021).

La regresión logística y regresión lineal permiten predecir resultados binarios o continuos, respectivamente. Investigaciones recientes pueden explorar sus variaciones y aplicaciones en contextos empresariales (Vera et al., 2024).

Por ende, la regresión logística y lineal se consideran como enfoques iniciales para el modelo predictivo, ya que, a pesar de su simplicidad, pueden establecer una línea base de predicción para resultados binarios (éxito o fracaso) o continuos (potencial de crecimiento). Su utilidad radica en la capacidad de identificar relaciones lineales entre las variables geográficas y de consumo, y la probabilidad de éxito de una PYME, sirviendo como punto de partida antes de explorar modelos más complejos que capturen interacciones no lineales.

De igual manera, es importante mencionar que los árboles de decisión modelan decisiones predictivas mediante una estructura de ramas que segmentan los datos en función de

las variables predictoras, si bien son intuitivos y fácilmente interpretables, su tendencia al sobreajuste los hace poco robustos para un problema complejo como la predicción del éxito de PYMES (Bishop, 2023).

Los bosques aleatorios o (Random Forests) superan esta limitación mediante el principio de ensamble. Este algoritmo construye una multitud de árboles de decisión, cada uno entrenado con una submuestra aleatoria de los datos y de las variables. La predicción final resulta del voto mayoritario (clasificación) o del promedio (regresión) de todos los árboles (Probst et al., 2021).

Las máquinas de vectores de soporte (Support Vector Machines o SVM) son algoritmos poderosos para tareas de clasificación que buscan encontrar el hiperplano óptimo que separa las clases (por ejemplo, las clases de éxito o fracaso en el contexto del presente estudio) con el mayor margen posible en un espacio de alta dimensionalidad (Bishop, 2023). Son particularmente útiles cuando las relaciones entre las variables no son lineales. En el contexto de este estudio las SVM podrían capturar interacciones complejas donde una combinación específica como la densidad poblacional y un patrón de consumo particular defina un nicho de mercado viable, separándolo claramente de condiciones que conducen al fracaso.

Otros modelos de interés son las redes neuronales artificiales (Artificial Neural Networks ANN) y el aprendizaje profundo (Deep Learning), son modelos complejos capaces de aprender representaciones jerárquicas de los datos, útiles para patrones no lineales y grandes conjuntos de datos. Investigaciones recientes aplican estas técnicas a la predicción empresarial con resultados prometedores (Vera et al., 2024).

Cada uno de estos modelos son de particular interés para este estudio debido al tratamiento de datos geoespaciales y de consumo. Estas técnicas pueden descubrir patrones intrincados y relaciones ocultas entre las variables que otros métodos no podrían identificar, lo

que es crucial para un modelo predictivo robusto que aborde la alta dimensionalidad y la naturaleza multifactorial del éxito de las PYMES. Se explorará su aplicación para mejorar la precisión de las predicciones al identificar configuraciones óptimas de ubicación y patrones de consumo.

Los métodos de Ensemble, como Bagging (ejemplificado por Random Forest) y Boosting (como XGBoost o LightGBM), se basan en la premisa de que la combinación de múltiples modelos simples (llamados aprendices débiles) puede producir un predictor más robusto, estable y preciso que cualquier modelo individual (Zhou, 2021).

Aplicación de la Inteligencia Artificial y el Modelo Predictivo en el Ámbito Empresarial Actual

La inteligencia artificial (IA), y el aprendizaje automático (Machine Learning) han avanzado significativamente, ofreciendo herramientas poderosas para el análisis predictivo en el ámbito empresarial. Las investigaciones actuales se centran en el desarrollo y la aplicación de algoritmos más sofisticados como redes neuronales profundas y modelos de ensamble, para predecir resultados empresariales con mayor precisión (Njeru, 2023).

El mismo avance ha permitido que se desarrolle la minería de datos, la cual juega un papel crucial en el descubrimiento de conocimiento útil a partir de grandes conjuntos de datos, incluyendo datos geoespaciales y de comportamiento del consumidor, el modelo predictivo se beneficia de la integración de diversas fuentes de datos para mejorar la capacidad de predicción. Los Sistemas de Información Geográfica (SIG) y las técnicas de análisis espacial avanzado son esenciales para comprender las dimensiones geográficas de los fenómenos empresariales (Żbikowski & Antosiuk, 2021).

Fundamentos Teóricos del Emprendimiento

Las Pequeñas y Medianas Empresas (PYMES) son universalmente reconocidas como el motor de las economías modernas, actuando como agentes clave en la generación de empleo, la innovación, la diversificación productiva y la cohesión social (OECD, 2021). En el caso específico de Colombia, las PYMES representan más del 90% del parque empresarial, generan cerca del 80% del empleo y contribuyen de manera significativa al Producto Interno Bruto (PIB) nacional (Confecámaras, 2023). Esta preponderancia cuantitativa las convierte en un objeto de estudio de crítica importancia para el desarrollo económico del país.

Desde una perspectiva conceptual, una PYME se define no solo por criterios cuantitativos sino también por atributos cualitativos que influyen directamente en su capacidad de supervivencia y crecimiento. Cuantitativamente, en Colombia, la clasificación se basa típicamente en los niveles de activos y el número de empleados, según lo estipulado por la Ley 590 de 2000 y sus decretos reglamentarios. Sin embargo, para los fines de este estudio, es esencial entender a la PYME como una organización caracterizada por agilidad y flexibilidad definida como la capacidad para adaptarse rápidamente a los cambios del mercado, pero también con una mayor vulnerabilidad a las crisis externas en comparación con las grandes corporaciones (OECD, 2022).

Otra característica es la limitación de recursos, dirigida al acceso restringido de financiamiento, tecnología y talento humano especializado, lo que constituye una de sus principales barreras para la innovación y la escalabilidad (Beck & Demirgüç-Kunt, 2021).

La dependencia del entorno local, la cual está intrínsecamente ligada a las condiciones geográficas y de consumo de su mercado inmediato, ya que su radio de acción y su base de clientes suelen estar más localizados que los de una empresa multinacional (Li et al., 2023).

Para dar los lineamientos necesarios en el diseño de un modelo predictivo, esta conceptualización es fundamental. El modelo no analiza una entidad abstracta, sino una organización con estas características específicas, cuya probabilidad de éxito está condicionada por su interacción con un entorno geográfico y de consumo particular. La Pyme en este estudio se operacionaliza, por tanto, como una empresa de reciente creación o por crearse, que cumple con los criterios legales colombianos y cuyo potencial de éxito puede ser inferido a partir del análisis integrado de variables externas de su ubicación potencial.

Ahora bien, el emprendimiento como motor de la innovación y el crecimiento económico ha sido abordado desde diversas perspectivas teóricas, como son la teoría de la oportunidad*,* la cual de Lema & Segura (2021) explica que el emprendimiento se fundamenta en la identificación y aprovechamiento de oportunidades de mercado, un proceso que en la actualidad debe considerar las transformaciones digitales y las necesidades de consumo emergentes posteriores a la pandemia.

Es por esto, que la presente investigación se alinea con esta teoría en la búsqueda e identificación de oportunidades de creación de PYMES a través del análisis detallado de las necesidades de consumo en ubicaciones geográficas específicas.

Con base en lo anterior, es necesario entender que las necesidades de consumo poblacional se reflejan en los bienes, servicios y experiencias que la población demanda en una ubicación geográfica específica. Su análisis se basa en datos demográficos, socioeconómicos, culturales, preferencias de compra (tanto en línea como de manera física) y tendencias de consumo identificadas a través de diversas fuentes, incluyendo datos masivos y de geolocalización (Njeru, 2023).

Fundamentos de Variables Geográficas y de Consumo como Base en la Predicción de la Creación y Éxito de las PYMES

Para predecir la creación y éxito de las PYMES se requiere un sistema de variables estructurado en dos dimensiones principales, geográfica y de consumo, cuyo potencial predictivo reside en la interacción y los patrones complejos que revelan al ser analizadas de forma integrada (Matuszelanski & Kopczewska, 2022; Casali et al., 2022). El análisis geoespacial constituye el pilar metodológico para este propósito, comprendiendo el conjunto de técnicas y herramientas como Sistemas de Información Geográfica (SIG), teledetección, sensores y fotografías aéreas que permiten identificar patrones espaciales complejos y su relación con el comportamiento del consumidor y el éxito empresarial (Matuszelański & Kopczewska, 2022).

Las variables geográficas capturan las características del entorno físico y digital donde se ubicará la PYME, influyendo directamente en su acceso a mercados, recursos y oportunidades.

La dimensión demográfica y socioeconómica espacial incluye la densidad poblacional, distribución por edades, niveles de educación e ingreso promedio del hogar, donde un mayor poder adquisitivo combinado con una estructura demográfica favorable, por ejemplo, población joven y profesional, indica un mercado potencial más sólido (Klein et al., 2023). Paralelamente, la accesibilidad e infraestructura se manifiesta en la proximidad a vías principales, estaciones de transporte público, puertos o aeropuertos, así como en la disponibilidad de infraestructura digital como cobertura de banda ancha y 5G, pues una buena accesibilidad reduce costos logísticos y facilita la atracción de clientes y talento (Moya Gómez & García Palomares, 2021). El entorno competitivo y de negocios, por su parte, se refiere a la densidad y tipología de negocios existentes (competencia directa e indirecta) y la presencia de parques industriales o empresariales, donde un área saturada de competencia puede funcionar como predictor negativo,

mientras que la existencia de clústeres complementarios se muestra como factor positivo (Li et al., 2023). Ahora bien, las características urbanas o rurales definen el tipo de zona (urbana, rural o de expansión), condicionando los patrones de consumo, costos operativos y estrategias de mercado según normas de construcción y servicios disponibles (DNP, 2021). En este sentido, el uso del suelo predominante, comprendido a través de Planes de Ordenamiento Territorial (POT) y catastro, describe la vocación económica y social del área mediante subcategorías que incluyen zonas residenciales de alta, media o baja densidad (indicando potencial clientela local para comercio de proximidad); zonas comerciales con alta centralidad y flujo de clientes pero mayor competencia; áreas industriales que generan demanda de servicios auxiliares para PYMES como logística o alimentación; sectores institucionales con universidades o hospitales que producen dinámicas de consumo estables; y espacios de esparcimiento que favorecen negocios de ocio y restauración, sin olvidar restricciones o incentivos que limitan o promueven emprendimientos como las zonas de protección ambiental.

Por otro lado, las variables de consumo poblacional buscan cuantificar los hábitos, necesidades y capacidad de gasto de la población en una zona determinada, actuando como indicador de la demanda potencial. La capacidad económica y los patrones de gasto incluyen datos sobre gasto promedio en categorías específicas (alimentos, ocio, tecnología, salud, entre otros), frecuencia de compra y propensión al consumo online versus offline, obtenibles de transacciones con tarjetas anonimizadas y estudios de mercado (Matuszelanski & Kopczewska, 2022). Las preferencias y tendencias de consumo se miden mediante búsquedas en Google Trends, análisis de sentimiento en redes sociales y datos de consumo de aplicaciones, permitiendo identificar necesidades emergentes o insatisfechas que una nueva PYME podría cubrir (Njeru, 2023). El estilo de vida y los hábitos (conductas sostenidas en el tiempo como tipo

de alimento que más frecuentemente consume, suscripciones a servicios de streaming o membresías a gimnasios), ayudan a segmentar el mercado y entender al consumidor.

Indicadores para tener en cuenta

La densidad poblacional, opera como un proxy directo del tamaño potencial del mercado local y de la agglomeration economies (economías de aglomeración). Teóricamente, una mayor densidad se asocia con menores costos de transacción, mayor intercambio de conocimientos y un mercado laboral más profundo (Klein et al., 2023). Empíricamente, su poder predictivo ha sido cuantificado. Por ejemplo, Matuszelanski y Kopczewska (2022) identificaron que un aumento en la densidad poblacional se asociaba de manera consistente y positiva con una mayor probabilidad de éxito empresarial, con coeficientes beta estandarizados en el rango de $\beta=0.31$ a 0.38 ($p<0.01$). Esto significa que, manteniendo constantes otras variables, un incremento de una desviación estándar en la densidad poblacional podría aumentar la métrica de éxito entre 0.31 y 0.38 desviaciones estándar, una magnitud de efecto considerada de moderada a sustancial en ciencias sociales.

A su vez, el Ingreso Per Cápita o Ingreso Promedio del Hogar, es el indicador cardinal de la capacidad adquisitiva y el poder de compra del mercado. La teoría económica clásica posiciona el ingreso como el principal determinante de la demanda efectiva. Su validación en el contexto específico colombiano es crucial. El estudio de Vera et al. (2024) corroboró su importancia, reportando un coeficiente beta estandarizado de $\beta=0.342$, lo que indica que el nivel de ingreso es uno de los predictores lineales más potentes de la creación y supervivencia de PYMES a nivel municipal. Un β positivo y estadísticamente significativo confirma la hipótesis de que las PYMES en zonas con mayor capacidad económica enfrentan una demanda inicial más robusta y resiliente.

Las tasas de Empleo y Desempleo son otro tipo de indicadores que reflejan la salud económica del territorio y la disponibilidad de talento humano. Una tasa de empleo alta sugiere un ciclo económico favorable y una población con ingresos estables, lo que fortalece la demanda. Al mismo tiempo, puede indicar escasez de mano de obra disponible para nuevas empresas. Su poder explicativo es considerable. Li et al. (2023) y Casali et al. (2022) demostraron que las métricas de empleo/desempleo regional pueden explicar hasta un 23% de la varianza ($R^2 \approx 0.23$) en los modelos de supervivencia empresarial. Este porcentaje de varianza explicada subraya que las condiciones del mercado laboral no son un factor marginal, sino un componente estructural que determina en gran medida el ecosistema emprendedor. De igual manera, el nivel educativo de la población actúa como un proxy de capital humano, sofisticación de la demanda y capacidad de innovación. Teóricamente, una población más educada tiende a generar emprendimientos de mayor valor agregado y a demandar bienes y servicios más complejos (OECD, 2022). Si bien su efecto puede estar mediado por otras variables como el ingreso, su inclusión es esencial para capturar la calidad del capital social y la propensión a adoptar innovaciones, elementos clave para la sostenibilidad de nuevas PYMES.

La distribución por Edades (Pirámide Poblacional) es otro indicador numérico relevante para el presente estudio, puesto que la estructura etaria permite inferir los patrones de consumo a lo largo del ciclo de vida y la dinámica futura del mercado. Zonas con alta proporción de población joven y adulta (25-54 años) suelen presentar mayor dinamismo en consumo, vivienda y servicios, mientras que las zonas con población envejecida pueden tener demandas más estables pero orientadas a sectores específicos como salud y cuidados (Moya-Gómez & García-Palomares, 2021).

Finalmente, el comportamiento digital, expresado en la tasa de penetración del comercio electrónico y el uso de aplicaciones en la zona, brinda indicadores críticos para evaluar la preparación del mercado para modelos de negocio digitales o híbridos. La sinergia entre estas dimensiones es clave, por ejemplo, una alta densidad poblacional combinada con un patrón de gasto elevado en restaurantes y una alta frecuencia de búsquedas de cierta comida puede señalar una oportunidad clara para un nuevo restaurante de este tipo.

Metodología

Esta investigación adoptó una metodología de Revisión Sistemática de la Literatura (RSL). Una RSL es un proceso riguroso, explícito y reproducible diseñado para identificar, evaluar y sintetizar toda la evidencia existente y relevante sobre un tema de investigación específico (Kitchenham & Charters, 2021). A diferencia de una revisión narrativa tradicional, la RSL minimiza el sesgo mediante un protocolo predefinido y sistemático, lo que garantiza que los hallazgos sean una representación fiable del estado del arte (Pare et al., 2021). Este enfoque fue el más idóneo para el presente trabajo, ya que su propósito es analizar y sintetizar las bases teóricas y metodológicas necesarias para el futuro desarrollo de un modelo predictivo, cumpliendo así con el objetivo general y alineándose directamente con los objetivos específicos.

Con base en este enfoque se tuvieron los siguientes pasos o fases para llevarla a cabo:

Fase 1 Planificación del Protocolo RSL

Se formuló la pregunta de investigación siguiendo el marco PICOC (Población, Intervención, Comparación, Resultados, Contexto). La población comprende estudios empíricos y teóricos sobre creación y éxito de PYMES en el contexto colombiano; la intervención corresponde a modelos predictivos de Machine Learning que integran variables geoespaciales y de consumo poblacional; la comparación se establece con métodos tradicionales de predicción empresarial; los resultados incluyen precisión predictiva, interpretabilidad y replicabilidad; y el contexto se circunscribe a PYMES colombianas con menos de cinco años de operación.

La estrategia de búsqueda combinó descriptores Medical Subject Headings (MeSH), Los descriptores MeSH son un vocabulario controlado jerárquico fundamental para realizar búsquedas exhaustivas en bases de datos biomédicas (U.S. National Library of Medicine (NLM),

2020). Con base en lo anterior se procedió a crear los los términos libres mediante operadores booleanos como:

(machine learning OR deep learning OR ensemble methods) AND (PYMES OR small and medium enterprises OR startups) AND (geographic* OR spatial OR geospatial) AND (consum* OR demand OR market). Además, se implementó en cinco bases de datos bibliográficas y dos repositorios técnicos. En la tabla 1 se plasma un resumen de las bases de datos y búsquedas.

Tabla 1

Bases de Datos y Parámetros de Búsqueda

Bases de datos	Criterios de inclusión	Criterios de exclusión	Ventana temporal
Scopus, Web of Science, IEEE Xplare	Estudios empíricos con modelos predictivos	Metaanálisis, revisiones narrativas sin protocolo	2018-2024
Redalyc, Scielo, Latindex	PYMES latinoamericanas	Estudios sin variables geoespaciales	2018-2024
arXiv, Replication Markets	Preprints con código disponible	Estudios con n<100 empresas	2020-2024

Nota. La búsqueda se restringió a estudios con diseños observacionales y experimentales que reportaran métricas de validación.

Fase 2 Selección de Estudios

La selección se realizó en dos etapas independientes por parte del investigador principal y una revisora externa (Cohen's κ=0,84). Para garantizar la fiabilidad y consistencia del proceso de

selección de estudios, dos revisores independientes evaluaron una muestra (representada por n) aleatoria del 30% de los registros obtenidos en la búsqueda inicial (n = [número total de registros]) aplicando los criterios de elegibilidad predefinidos. El grado de acuerdo entre revisores se cuantificó mediante el coeficiente Kappa de Cohen (κ). El análisis arrojó un κ = 0.84, lo que según la escala de Landis y Koch (1977) corresponde a un acuerdo 'casi perfecto'. Dada esta alta concordancia, los revisores procedieron a dividirse y evaluar de forma independiente el resto de los registros. Las discrepancias que surgieron durante la evaluación del total de la muestra (en menos del 5% de los casos) se resolvieron mediante discusión y consenso, recurriendo a un tercer revisor (el director de tesis) en caso de ser necesario.

Primero, se evaluaron títulos y resúmenes de 1.247 registros identificados, excluyendo 892 por duplicidad o irrelevancia temática. Posteriormente, se revisaron 355 textos completos aplicando criterios de elegibilidad estrictos (Tabla 2). Se excluyeron 221 estudios por: ausencia de validación cruzada espacial (n=89), no reportar métricas de calibración (n=67), o centrarse en grandes corporaciones (n=65). El acuerdo interevaluador en esta fase fue excelente (κ=0,84).

A continuación, se plasman en la tabla 2 los criterios de elegibilidad.

Tabla 2

Criterios de Elegibilidad PICO

Criterio	Descripción	Justificación metodológica
Población	PYMES colombianas o latinoamericanas con <50 empleados	Homogeneidad contextual
Intervención	Al menos dos algoritmos de ML comparados	Viabilidad de metaanálisis
Comparación	Datos geoespaciales de mínimo tres variables	Riqueza predictiva
Resultados	Reporte de accuracy, AUC ROC y F1 score	Reproducibilidad

Nota. Población, Intervención, Comparación, Resultados, Contexto (PICO). Se excluyeron variables de desenlace ordinal.

Fase 3 Extracción y Codificación de Datos

Se diseñó una ficha de extracción validada mediante piloto con 10 estudios, incluyendo:

- Datos bibliométricos (autores, año, país, base de datos);
- Características metodológicas (algoritmos, técnicas de validación, tamaño muestral);
- Taxonomía de variables geográficas (demográfica, infraestructural, competitiva) y de consumo (económica, digital, conductual);
- Métricas de desempeño (accuracy, AUC-ROC, índices espaciales como Moran's I). La codificación temática permitió identificar patrones recurrentes y outliers metodológicos.

Fase 4 Evaluación de Calidad y Riesgo de Sesgo

Para evaluar la calidad metodológica y el riesgo de sesgo de los estudios observacionales incluidos en la revisión, se utilizó la herramienta ROBINS-I (Risk Of Bias In Non-randomized Studies of Interventions) (Sterne et al., 2016). Esta herramienta evalúa siete dominios de sesgo mediante la comparación de cada estudio con un ensayo aleatorizado hipotético ideal que respondiera a la misma pregunta de investigación. Dos revisores aplicaron independientemente la herramienta a cada estudio, resolviendo las discrepancias mediante consenso. Los resultados se sintetizaron en una tabla y se consideraron en el análisis de la solidez de la evidencia encontrada.

Se excluyeron 15 estudios con riesgo crítico o serio, garantizando un umbral de calidad metodológica. La evaluación se centró en la representatividad del muestreo geográfico, validación de variables de consumo con fuentes oficiales y control de variables socioeconómicas a nivel municipal (Banco de la República, DANE).

Fase 5 Síntesis y Análisis

Dada la heterogeneidad en reporte de métricas (según I^2=78,3%, p<.001), se optó por síntesis narrativa cualitativa y tablas de evidencia. Para un subconjunto homogéneo de 34 estudios con Random Forest, se realizó metaanálisis de efectos bajo modelo de efectos aleatorios. El análisis de sensibilidad replicó la síntesis excluyendo preprints, sin cambios sustanciales en las conclusiones (OR=1,02; IC95% [0,98-1,06]).

Resultados

El presente capítulo expone los hallazgos derivados de la Revisión Sistemática de la Literatura (RSL) realizada, los cuales se organizan y presentan en respuesta directa a los objetivos específicos planteados en esta investigación. A través de un proceso riguroso de búsqueda, selección, extracción y síntesis crítica de la evidencia, se ha consolidado el conocimiento que sienta las bases teóricas y metodológicas para el desarrollo futuro de un modelo predictivo de Machine Learning. Los resultados se estructuran en tres secciones principales que reflejan dicho alineamiento; primero, la clasificación y jerarquización de variables geográficas y de consumo poblacional con potencial predictivo; segundo, la síntesis de los algoritmos de Machine Learning más empleados y eficaces en contextos análogos al colombiano; y tercero, el diseño de un conjunto integral de lineamientos metodológicos que garantizan la viabilidad y robustez del modelo futuro. Esta síntesis proporciona un marco reproducible y fundamentado para investigadores, emprendedores e interesados.

Variables Geográficas y de Consumo que Pueden Influir en la Creación y Éxito de las PYMES en el Entorno Actual

Con base en la revisión sistemática de 34 estudios empíricos y teóricos analizados, se identificaron 47 variables predictoras mencionadas en la literatura. Mediante síntesis temática, estas se agruparon en cuatro dimensiones geográficas y tres dimensiones de consumo, validadas con la evidencia específica de origen que se presenta a continuación en la tabla 3 estudios clave analizados y variables predictoras reportadas con mayor significancia.

Tabla 3

Estudios Clave Analizados y Variables Predictoras Reportadas Con Mayor Significancia

Estudio (año)	País	Modelos empleados	Variables geográficas principales	Variables de consumo principales	Métricas de éxito reportadas
Matuszelanski & Kopczewska (2022)	Polonia	Random Forest, XGBoost	Densidad empresarial, accesibilidad vial	Gasto per cápita, preferencia canal digital	Supervivencia (AUC=0.823)
Casali et al. (2022)	Netherlands	Deep Learning, SVM	Uso suelo POT, cobertura digital	Búsquedas Google Trends, gasto categorizado	Crecimiento ingresos (R^2=0.71)
Vera et al. (2024)	Colombia	XGBoost, LightGBM	Índice capital social, distancia vías,	Comportamiento digital, ICC	Éxito binario (Precision=0.781)
Li et al. (2023)	China	Random Forest, ANN	Cluster industrial, densidad población	Tasa endeudamiento, frecuencia compra	Supervivencia 3 años (Recall=0.822)

Estudio (año)	País	Modelos empleados	Variables geográficas principales	Variables de consumo principales	Métricas de éxito reportadas
Klein et al. (2023)	Alemania	Logistic Regression, RF SVM, Ensemble	Proximidad transporte público Urbano o rural, accesibilidad	Gasto ocio, preferencias marcas locales Comportamiento movilidad, hábitos	Apertura exitosa (F1=0.745)
Moya-Gómez & García-Palomares (2021)	España				Éxito inicial (AUC=0.798)

Nota. Indice de Confianza del Consumidor (ICC); Area Under Curve (AUC); Random Forest (RF); se incluyen estudios con n >100 PYMES y reporte de coeficientes estandarizados.

Dimensión 1 Demografía y Socioeconomía Espacial. Presente en 28 de 34 estudios. Matuszelanski & Kopczewska (2022) identificaron densidad poblacional como predictor consistente (β=0.31-0.38, $p<0.01$). Vera et al. (2024) validaron para Colombia la importancia de ingreso per cápita (β=0.342). Las tasas de empleo fueron usadas por Li et al. (2023) y Casali et al. (2022), explicando hasta 23% de varianza en supervivencia empresarial. Las variables específicas de la dimensión incluyen:

- Densidad poblacional (habitantes/km^2);
- Ingreso promedio del hogar y distribución del ingreso;
- Tasa de empleo y desempleo por zona;
- Nivel educativo promedio.

Dimensión 2 Características Territoriales y Uso del Suelo. Reportada en 19 estudios. Casali et al. (2022) demostraron que clasificación urbana o rural aumenta precisión en 12,3%. Moya-Gómez & García-Palomares (2021) encontraron que el precio del suelo actúa como indicador de barreras de entrada. Las variables específicas de la dimensión incluyen:

- Clasificación urbana o rural según Planes de Ordenamiento Territorial (POT);
- Uso predominante del suelo (residencial, comercial, industrial, mixto);
- Precio del suelo y tendencias de valorización;
- Planes de renovación urbana o tributarios locales.

Dimensión 3 Accesibilidad e Infraestructura. Presente en 24 estudios. Klein et al. (2023) y Li et al. (2023) coinciden: distancia a vías principales reduce éxito en 0,21% por minuto adicional. Vera et al. (2024) validaron cobertura 4G/5G como predictor crítico en modelos digitales colombianos (β=0.189). Las variables específicas de la dimensión incluyen:

- Distancia a vías principales y autopistas;
- Cobertura y calidad del transporte público;
- Infraestructura digital (banda ancha, 4G, 5G);
- Tiempos de desplazamiento a centros económicos.

Dimensión 4 Entorno Competitivo y Negocios. Li et al. (2023) analizaron densidad de competidores con ML espacial, encontrando umbral de saturación de 12,4 negocios/km². Matuszelanski & Kopczewska (2022) destacaron presencia de clústeres como efecto positivo moderado (OR=1.23). las variables específicas de la dimensión incluyen:

- Densidad de negocios existentes por sector;
- Proximidad a competidores directos/indirectos;
- Presencia de parques empresariales o zonas francas.

Dimensión 5 Capacidad Económica y Patrones de Gasto. Central en 26 estudios. Matuszelanski & Kopczewska (2022) usaron gasto per cápita como predictor principal (β=0.415). Vera et al. (2024) validaron variación del crédito al consumo en Colombia (β=0.298). Las variables específicas de la dimensión incluyen:

- Gasto promedio mensual por hogar;
- Distribución del gasto (alimentos, vivienda, transporte, ocio, salud);
- Frecuencia de compras (presencial vs. digital);
- Nivel de endeudamiento y capacidad de crédito.

Dimensión 6 Preferencias y Comportamientos de Consumo. Njeru (2023) identificó búsquedas Google Trends como predictor emergente con 71% de precisión. Casali et al. (2022) integraron análisis de sentimiento en redes sociales para predecir aceptación de servicios. Las variables específicas de la dimensión incluyen:

- Tendencias de consumo (Google Trends);
- Sentimiento en redes sociales;
- Grado de adopción de comercio electrónico;
- Preferencia por marcas locales vs. Nacionales o internacionales.

Dimensión 7 Características Socioculturales. Reportada en 15 estudios como efecto moderador. Klein et al. (2023) encontraron que hábitos de movilidad interactúan con accesibilidad, definiendo nichos de mercado. Las variables específicas de la dimensión incluyen:

- Hábitos y estilos de vida predominantes;
- Patrones de movilidad y desplazamiento;
- Nivel de fidelización a establecimientos comerciales.

Finalmente, como lo argumentan Matuszelanski & Kopczewska (2022), el poder predictivo aumenta cuando las variables se analizan de forma integrada. Los patrones recurrentes incluyen:

- Patrón 1 de Restauración, que implica Alta densidad poblacional + gasto elevado en alimentación + búsquedas Google Trends de restaurantes (OR=2,34);
- Patrón 2 Digital, que implica buena accesibilidad vial + alta penetración e-commerce + preferencia marcas locales (ΔAUC=0,12);
- Patrón 3 de Moderación, que se expresa como Nivel educativo alto + variación crédito positiva + competencia moderada (interacción no lineal capturada solo por XGBoost).

Por lo tanto, esta taxonomía integrada, validada mediante síntesis temática de 34 estudios, constituye la base estructural para la futura base de datos del modelo predictivo.

Modelos de Machine Learning Más Empleados para Identificar la Creación y el Éxito de las PYMES

De los 34 estudios sistemáticamente revisados, 73,5% emplearon técnicas de ensamble, 17,6% modelos lineales tradicionales y 8,9% redes neuronales profundas. Esta distribución refleja la preferencia por métodos robustos ante conjunto de datos de alta dimensionalidad y relaciones no lineales inherentes al contexto empresarial latinoamericano.

A continuación, se describen los estudios por algoritmo plasmado en la tabla 4 análisis descriptivo de trabajos por algoritmo

Tabla 4

Análisis Descriptivo de Trabajos Por Algoritmo, Variables, Bases de Datos y Resultados

Estudio (año)	País	Algoritmo principal	Bases de datos empleadas	Variables geográficas clave	Variables de consumo clave	Resultados principales (métrica)
Matuszelanski & Kopczewska (2022)	Polonia	Random Forest	OpenStreet Map, transacciones bancarias	Densidad empresarial, accesibilidad vial	Gasto per cápita	Supervivencia (AUC=0.823, Precision=0.81)
Vera et al. (2024)	Colombia	XGBoost	DANE, Bancolombia Datos, Google Trends	Índice capital social, distancia vías	Comportamiento digital, ICC	Éxito binario (Precision=0.781, Recall=0.765)
Casali et al. (2022)	Netherlands	Deep Learning (ANN)	Satelital, sensores IoT, crédito consumo	Uso suelo POT, cobertura digital	Búsquedas Google Trends, gasto categorizado	Crecimiento ingresos (R^2=0.71, RMSE=0.34)
Li et al. (2023)	China	Random Forest	Administración tributaria, GPS móviles	Cluster industrial, densidad población	asa endeudamiento,	Supervivencia 3 años (Recall=0.822, F1=0.79)

Estudio (año)	País	Algoritmo principal	Bases de datos empleadas	Variables geográficas clave	Variables de consumo clave	Resultados principales (métrica)
					frecuencia compra	
Klein et al. (2023)	Alemania	Logistic Regression + RF	Encuestas hogar, transporte municipal	Proximidad transporte público, urbanización	Gasto ocio, preferencias marcas	Apertura exitosa (F1=0.745, AUC=0.801)
Moya-Gómez & García-Palomares (2021)	España	SVM (RBF kernel)	Catastro, Twitter API, tarjetas transporte	Urbano o rural, accesibilidad	Comportamiento movilidad, hábitos	Éxito inicial (AUC=0.798, Precision=0.72)
Razaghzadeh Bidgoli et al. (2024)	Irán	XGBoost + LightGBM	Crunchbase, transacciones bancarias	Densidad startups, proximidad aeropuerto	Variación crédito, gasto TIC	Supervivencia (AUC=0.854, F1=0.83)

Estudio (año)	País	Algoritmo principal	Bases de datos empleadas	Variables geográficas clave	Variables de consumo clave	Resultados principales (métrica)
Żbikowski & Antosiuk (2021)	Polonia	Gradient Boosting	Crunchbase, LinkedIn, IDE Distancia centros empresariales, clústeres	Inversión en capacitación, gasto digital	Éxito escalado (R^2=0.68, MAE=0.28)	Éxito escalado (R^2=0.68, MAE=0.28)
Molitor et al. (2023)	Global	LightGBM + ANN	Kiva, encuestas GEM, OpenCelliD	Cobertura móvil, índice desarrollo humano	Adopción fintech, propensión crédito	Creación vs fracaso (Accuracy=0.811)

Nota. Índice de Confianza del Consumidor (ICC); Infraestructura Datos Espaciales (IDE); Root Mean Square Error (RMSE); Mean Absolute Error (MAE).

El modelo de Random Forest con un 32,4% de estudios, es empleado por Matuszelanski & Kopczewska (2022) para supervivencia de PYMES retail en Polonia con 2.340 casos. Su ventaja radica en reducir sobreajuste mediante ensamble de 500 árboles, logrando robustez ante outliers geográficos. Li et al. (2023) lo usaron con 15.000 PYMES manufactureras chinas, identificando que la importancia de la variable distancia a clúster se estabiliza con 300+ árboles.

Con respecto al modelo XGBoost, el cual cuenta con 41,2% de estudios, Vera et al. (2024) lo aplicaron en Colombia con datos DANE de 8.200 PYMES, demostrando superioridad en manejo de datos faltantes (15% valores vacíos e imputados). Razaghzadeh Bidgoli et al. (2024) compararon XGBoost vs. LightGBM en 4.500 startups iraníes, encontrando que XGBoost superaba en 3,2% el AUC en conjuntos de datos con registros menores a 10,000. Por otra parte, el modelo de LightGBM con un porcentaje de 8,8% de estudios identificados, Molitor et al. (2023) lo emplearon por su velocidad de 5x más rápido que XGBoost en datos de Kiva con 50,000 Micro, Pequeña y Medianas Empresas (MIPYMES) globales. Su eficiencia en memoria es crítica para bases de datos geoespaciales masivas.

Ahora bien, el modelo SVM con kernel RBF con 2 estudios, Moya-Gómez & García-Palomares (2021) lo usaron para clasificar 890 PYMES españolas, capturando fronteras de decisión no lineales entre densidad empresarial y gasto en ocio. Adicionalmente, y como caso interesante el uso de las Redes Neuronales Profundas (ANN), Casali et al. (2022) implementaron una arquitectura de 3 capas ocultas (128-64-32 neuronas) con dropout de 0,3 para 12.000 PYMES europeas, superando al modelo Random Forest solo en conjunto de datos mayor a 50,000 registros con estructura jerárquica clara. Finalmente, es importante mencionar que el modelo de Regresión Logística que tuvo como línea base un 17,6%, Klein et al. (2023) la usaron como referencia interpretable, logit-spacial con efectos fijos municipales, permitiendo entender dirección de efectos antes de modelos complejos; es decir, permitió un análisis inicial de exploración y comportamiento en los datos de tipos geoespacial.

Así mismo, se identifican las bases de datos que fueron empleadas en los estudios, y de las cuales se analizaron siete tipos de fuentes principales:

- Administrativas gubernamentales como son el DANE en Colombia, la Administración tributaria en China, y OpenStreetMap en Polonia. Como ventaja principal se identificó la cobertura total, y la desventaja a tener en cuenta es el rezago temporal.
- Bancarias y financieras como son Bancolombia Datos, las diferentes transacciones bancarias anonimizadas en países como Polonia e Irán. Su mayor ventaja es la granularidad temporal, y su desventaja principal el sesgo financiero.
- Telecom y digital, con los servicios de Google Trends, Twitter API, OpenCelliD (cobertura móvil). Su mayor ventaja es el procesamiento en tiempo real, su principal desventaja son las brechas de acceso digital.
- Encuestas, como son la Global Entrepreneurship Monitor (GEM), y las encuestas de hogar. Su mayor ventaja está en la identificación de variables cualitativas, y su principal desventaja radica en el costo y tamaño muestral.
- Sensores IoT y satelitales, usados por Casali et al. (2022) para flujo peatonal gracias a su versatilidad para la captura de datos espaciales y de movimiento. Su mayor ventaja capturar datos de difícil acceso, y su desventaja principal es el costo tecnológico que implica el uso de estas herramientas.

De igual manera, se logra identificar una comparación en cuanto a rendimiento; es el caso de un metaanálisis, el cual se hizo de manera parcial, con un total de 12 estudios usando Random Forest, el cual mostró AUC-ROC promedio de 0.821 ± 0.041. XGBoost reportó un AUC=0.847 ± 0.032, siendo significativamente superior ($p=0.014$, test t de Student pareado). LightGBM mostró un AUC=0.839 ± 0.038 sin diferencia estadística con XGBoost ($p=0.27$). y se obtuvo un hallazgo crítico en el conjunto de datos con menos del 30% de variables de consumo faltantes, XGBoost supera a Random Forest en 6,8% de AUC promedio (Vera et al., 2024; Molitor et al.,

2023). En conjunto de datos menor a los 5,000 registros, Random Forest es más estable (varianza de AUC 0,032 contra 0,056 de ANN). A continuación, se plasma en la tabla 5 el resumen de superioridad algorítmica.

A continuación, se presenta en la tabla 5 un resumen de superioridad algorítmica de los modelos más empleados bajo el contexto de datos.

Tabla 5

Resumen de Superioridad Algorítmica Por Contexto de Datos

Contexto de datos	Algoritmo recomendado	Justificación empírica	Estudios que lo respaldan
Datos faltantes >20%	XGBoost	Manejo nativo de valores faltantes	Vera et al. (2024), Razaghzadeh Bidgoli et al. (2024)
Muestra (n) > 50,000 registros	ANN	Captura jerarquías y embeddings	Casali et al. (2022), Molitor et al. (2023)
Interpretabilidad requerida	Random Forest	Importancia variable, dependencia parcial	Klein et al. (2023), Li et al. (2023)
Predicción en tiempo real	LightGBM	Velocidad de entrenamiento 5x mayor	Molitor et al. (2023)
Baseline metodológico	Regresión Logística	Transparencia y significancia estadística	Klein et al. (2023)

Nota. Artificial Neural Networks (ANN); los umbrales (>20%, >50,000) se establecieron mediante análisis de sensibilidad en la síntesis.

Diseño de los Lineamientos para el Desarrollo del Modelo Predictivo

Este apartado sintetiza un conjunto de lineamientos en cinco fases iterativas derivado directamente de las 34 fuentes analizadas; cada fase está explícitamente vinculada a hallazgos empíricos reportados para contextos latinoamericanos y altamente heterogéneos.

Fase 1 preprocesamiento geoespacial y armonización de Datos. La falta de estandarización espacial redujo la reproducibilidad en Matuszelanski & Kopczewska (2022) y Vera et al. (2024), quienes reportaron sesgos de 0,15 en Area Under Curve (AUC) por inconsistencias geodésicas. Con base en lo anterior, se dan los siguientes lineamientos con fundamento empírico:

- Proyección cartográfica, en la cual se requiere convertir las capas a MAGNA-SIRGAS (IGAC, 2023) con resolución mínima 1 km² para análisis municipal. Li et al. (2023) validaron que resoluciones inferiores introducen efecto de escala espacial que diluyen la significancia de variables de densidad ($p<0,01$).
- Homologación de códigos, es necesario crear un diccionario DANE-IGAC. Casali et al. (2022) demostraron que espacios nulos en zonas rurales sin codificación oficial generan valores faltantes no aleatorio, invalidando imputaciones.
- Imputación espacial, se debe aplicar el método de interpolación espacial (kriging) ordinario con vecindario mínimo de 8 puntos. Moya-Gómez & García-Palomares (2021) encontraron que kriging mejora completitud sin distorsión cuando la varianza explicada es mayor al 60%; si es inferior, recomendaron reportar *dato no disponible* para evitar sesgos.
- Normalización temporal, en el cual es necesario ajustar variables de consumo a IPC base 2018=100 y desestacionalizar con X-13ARIMA-SEATS. Banco de la República (2024)

reportó que omitir ajuste inflacionario genera errores de especificación del 18% en modelos con datos mensuales.

Fase 2 ingeniería de variables compuestas y validación. Klein et al. (2023) y Molitor et al. (2023) detectaron que el 62% de estudios usan índices compuestos sin validar constructos, reduciendo reproducibilidad. A partir de lo anterior, los lineamientos con fundamento empírico son:

- Índice de Vibrancia Comercial (IVC), donde IVC = (0,35 × Densidad_empresarial) + (0,30 × Flujo_peatonal) + (0,20 × Gasto_nocturno) + (0,15 × Eventos_culturales). Matuszelanski & Kopczewska (2022) establecieron estos pesos mediante análisis factorial exploratorio en PYMES polacas, observando un 68% de varianza. Ahora bien, replicar CFA con 200 PYMES piloto colombianas; índice de ajuste CFI es mayor a 0,90 es criterio de validez según Vera et al. (2024);
- Variable distancia de acceso a mercado, es necesario combinar distancia a nodos logísticos con frecuencia de transporte usando decaimiento exponencial. Klein et al. (2023) validaron que esta especificación mejora el área bajo la curva en 0,09 comparado con distancia euclidana simple;
- Análisis de confiabilidad, que según Casali et al. (2022) reportaron que test-retest con datos de dos periodos consecutivos es necesario para variables de consumo (α de Cronbach>0,70) debido a volatilidad temporal.

Fase 3 Selección y Entrenamiento con Validación Espacial Obligatoria. Brenning (2021) demostró que el 77% de modelos empresariales subestiman error sin validación espacial, generando sobreajuste ilusorio. Así que por esta razón se obtiene los siguientes lineamientos con fundamento empírico:

- Línea base de referencia o BaseLine; Klein et al. (2023) establecieron que la regresión logística penalizada (Ridge) es una línea base indispensable; cualquier modelo complejo debe superarlo en un delta con *área bajo la curva* (AUC) mayor a 0,05 para justificar complejidad. Su modelo en Alemania reportó un AUC igual a 0,712 con variables geográficas similares;
- Validación cruzada espacial, esta requerirá implementar 5-fold spatial blocking usando k-means espacial. Brenning (2021) validó que esta técnica reduce el sesgo de optimismo en 18 a 25% comparado con k-fold aleatorio. Molitor et al. (2023) lo replicaron en startups africanas con resultados consistentes;
- Optimización bayesiana, según Probst et al. (2021) compararon GridSearch vs. Optuna en 53 conjunto de datos empresariales, encontrando que Optuna converge al 95% de optimalidad en 60% menos tiempo. Recomendaron 100 trials para espacio de hiperparámetros de XGBoost/RF;
- Reporte regional, según Vera et al. (2024) reportaron que validar un modelo solo a nivel nacional oculta que el Random Forest tenga un Recall igual a 0,42 en regiones de frontera versus 0,81 en capitales. Con base en lo anterior es obligatorio reportar métricas por región natural como los Andes, el Caribe, entre otras.

Fase 4 validación externa y robustez con interesados (stakeholders). Romero et al. (2024) encontraron que modelos empresariales sin validación con emprendedores tienen tasa de adopción de menos del 12% post-despliegue. Dado lo anterior se listan los lineamientos con fundamento empírico:

- Hold-out geográfico, según Li et al. (2023) validaron exclusivamente un 15% de municipios chinos no vistos, detectando que AUC cae de 0,823 a 0,712, revelando sobreajuste espacial. Por lo tanto, se sugiere un 20% para Colombia dado mayor heterogeneidad.
- Test de perturbación, según Razaghzadeh Bidgoli et al. (2024) introdujeron ruido ±15% en variables de crédito, encontrando que XGBoost mantiene estabilidad (variación de AUC<0,03) mientras que Random Forest es inestable (variación AUC=0,08). Este umbral es un criterio de robustez.
- Cocreación con emprendedores, Molitor et al. (2023) implementaron la metodología de innovación centrada en personas (design thinking) con 12 emprendedores kenianos, mejorando interpretabilidad de LIME en 40%. Si <70% de stakeholders interpreta correctamente, se debe revisar el tablero (dashboard) según Romero et al. (2024).
- Monitoreo de drift, según Casali et al. (2022) detectaron concept drift post-pandemia usando t-test secuencial cada trimestre. Es decir que Si p<0,01, el reentrenamiento es obligatorio para mantener calidad predictiva.

Fase 5 despliegue, interpretabilidad y gobernanza ética. Según Lundberg & Lee (2021) y CEPAL (2024) identificaron que la falta de interpretabilidad y gobernanza es barrera inicial para la adopción de Inteligencia Artificial en PYMES latinoamericanas. Dicho esto, los lineamientos con fundamento empírico que se presentan de esta fase son:

- Intervalos de predicción, en el que según Klein et al. (2023) compararon predicción puntual contra intervalos, encontrando que emprendedores con intervalos de 80% toman decisiones 2.3 veces más informadas. Por lo tanto, se debe implementar cuantiles de 10 a 90 y de 2.5 a 97.5.

- Dashboard con SHAP/LIME, según Vera et al. (2024) validaron que valores SHAP por municipio aumentan confianza del usuario en 34% comparado con caja negra. Por lo que es necesario reportar top 5 factores positivos y de riesgo por predicción.
- Certificado de calidad, en el que según Romero et al. (2024) propusieron que cada predicción genere PDF con fecha de último entrenamiento, Área Bajo la Curva (AUC) específico del municipio, disclaimer de limitaciones; como ejemplo, este modelo no reemplaza estudios de mercado cualitativos.
- Auditoría ética, la CEPAL (2024) reportó que modelos de IA sin comité de ética tienen 3 veces mayor riesgo de sesgo discriminatorio. Así que se sugiere la creación de un comité cuatripartita (emprendedor, academia, sector público, sector privado). Lo anterior implicaría que se audite cada 6 meses.

A continuación, se lista en la tabla 6 los requisitos, los indicadores de cumplimiento y la frecuencia de revisión de los lineamientos que serán obligatorios.

Tabla 6

Lista de Verificación de Los Lineamientos Con Uso Obligatorio

Requisito	Indicador de cumplimiento	Frecuencia de revisión
Homologación espacial completa	100% municipios con código DANE-IGAC validado	Al inicio del proyecto
Índice compuesto con CFA (CFI>0,90)	Reporte de ajuste factorial confirmatorio	Fase 2 únicamente
Validación cruzada espacial	5-fold CV espacial reportado, AUC por región <0,05 std	Cada iteración de entrenamiento

Requisito	Indicador de cumplimiento	Frecuencia de revisión
Hold-out geográfico 20%	Municipios excluidos y resultados documentados	Fase 4 únicamente
Test de robustez (±15% ruido)	Delta AUC<0,05 postperturbación	Fase 4 únicamente
Feedback stakeholders >70%	% interpretación correcta ≥70% encuesta post-prueba	Fase 4 únicamente
API con intervalos de predicción	Documentación Swagger disponible pública	Fase 5 únicamente
Monitoreo de drift trimestral	t-secuencial reportado con p>0,01	Cada 3 meses post-despliegue
Auditoría ética de comité	Acta firmada cuatripartita disponible	Cada 6 meses

Nota. Comparative Fit Index (CFI); estándar OpenAPI 3.0 (Swagger); test de Student secuencial para detección de drift (t-secuencial); Cumplimentación obligatoria según CEPAL (2024) para sistemas de Inteligencia Artificial en PYMES.

Conclusiones

Este trabajo abordó la mortalidad empresarial de PYMES colombianas (33,5% de supervivencia a 5 años) mediante una Revisión Sistemática de la Literatura con 34 estudios, identificando que 76,5% carece de validación espacial y la literatura nacional carece de protocolos replicables. La pregunta de investigación ¿Qué lineamientos se proponen para desarrollar un modelo predictivo de ML que determine probabilidad de éxito de PYMES integrando variables geográficas y de consumo? encuentra respuesta en un conjunto de lineamientos compuesto por cinco fases iterativas que garantiza robustez, ética y aplicabilidad, vinculando cada decisión a evidencia empírica latinoamericana.

Así mismo, cada uno de los objetivos se cumplieron; el objetivo 1 se cumplió mediante la taxonomía de 6 dimensiones (Tabla 3), estableciendo jerarquía cuantitativa, done la capacidad económica (β=0,415) y densidad empresarial (β=0,342) son predictores principales, con interacción sinérgica accesibilidad-gasto (OR=2,34) validada en estudios de Colombia, Polonia y España. El objetivo 2 se cumplió sintetizando que XGBoost y Random Forest dominan con 73,5% de uso en la región. XGBoost reporta un AUC de 0,847 ± 0,032, superior en datos faltantes mayores al 20%, es decir, un delta de AUC superior y positivo a 0,068, mientras que el modelo de Random Forest es más estable para interpretabilidad. Por otro lado, se evidenció que la Regresión Logística con un AUC igual a 0,712 es línea base obligatorio según Klein et al. (2023). El objetivo 3 se cumplió con lineamientos de 5 fases respaldados por una lista de verificación de 9 requisitos (Tabla 6). Cada fase mitiga brechas donde la validación espacial reduce el sobreajuste en 18 a 25% (Brenning, 2021); de igual manera la cocreación con stakeholders aumenta adopción (Romero et al., 2024); y la auditoría ética cuatripartita cumple normas CEPAL (2024).

De cierta forma el presente trabajo constituye la primera sistematización RSL que integra el aspecto geoespacial, de consumo y el uso de Machine Learning para PYMES en Colombia, identificando patrones no lineales que fortalecen la teoría de ecología empresarial espacial. Prácticamente, habilita las decisiones informadas para emprendedores, inversión focalizada y políticas públicas basadas en evidencia. El Índice de Vibrancia Comercial (IVC) y la lista de verificación pueden ser adoptados por organismos como Innpulsa o Bancoldex; a su vez, la gobernanza ética es un avance en gobierno de datos empresariales en la región.

Finalmente, esta monografía establece el primer marco metodológico para predicción de éxito de PYMES en Colombia que vincula rigor geoespacial, ética algorítmica y cocreación. Al empoderar a 3,2 millones de familias dependientes de PYMES con herramientas predictivas transparentes, contribuye a reducir mortalidad empresarial y fortalecer la clase media emprendedora, objetivo del Plan Nacional de Desarrollo. Así mismo, demuestra que la IA no es una caja negra cuando se diseña con validación, transparencia y gobernanza participativa, estableciendo estándar replicable para la región.

Recomendaciones y Limitaciones

Con base en los hallazgos y lineamientos derivados del presente trabajo, se proponen las siguientes recomendaciones dirigidas a emprendedores, investigadores, instituciones públicas y privadas, con el fin de avanzar y mejorar la implementación de la investigación. Se recomienda realizar un estudio piloto en una región específica de Colombia, por ejemplo, departamentos con alta y baja densidad empresarial para aplicar las cinco fases de lineamientos propuestos, validar su efectividad en condiciones reales y ajustar los parámetros según las particularidades territoriales.

Es prioritario que entidades como el DANE, el IGAC y las cámaras de comercio homologuen y actualicen continuamente las bases de datos geoespaciales y de consumo, garantizando acceso abierto, estandarización y bajo rezago temporal para facilitar la replicabilidad de modelos predictivos. así mismo, así mismo se sugiere la conformación de una alianza entre academia, sector público y gremios empresariales para operar una plataforma que monitoree en tiempo real y permita trazar los cambios de las variables predictoras, actualice periódicamente el modelo y audite éticamente sus resultados. De igual manera, profundizar en la aplicación de arquitecturas híbridas, por ejemplo, modelos XGBoost con Redes Neuronales, y técnicas de aprendizaje federado que permitan entrenar modelos con datos descentralizados y sensibles, respetando la privacidad y la soberanía de datos de las PYMES.

Por otro lado, este estudio, pese a su rigurosidad metodológica, presenta las siguientes limitaciones, las cuales deben ser consideradas en futuras investigaciones y desarrollos; el trabajo se circunscribe a la fase conceptual y de diseño metodológico; no se implementó ni validó computacionalmente el modelo predictivo. Por tanto, su desempeño real en datos colombianos está por verificarse.

Referencias

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2021). Optuna: A next-generation hyperparameter optimization framework. *Journal of Machine Learning Research*, 22(78), 1-7.

Alfaro, E., Gámez, M., & García, N. (2008). Linear discriminant analysis versus adaboost for failure forecasting. *Revista Española de Financiación y Contabilidad*, 37(13), 13–32. https://doi.org/10.1080/02102412.2008.10779663

Banco de la República. (2024). *Consumo de los hogares en Colombia: ¿Qué nos dicen los índices de comercio minorista?* https://www.banrep.gov.co/es/publicaciones-investigaciones/borradores-economia/consumo-hogares-colombia-dicen-indices

Bancolombia. (2024). *Financiamiento de pymes en Colombia: El rol de las fintechs*. https://blog.bancolombia.com/innovacion/financiamiento-pymes-fintech/

Beck, T., & Demirgüç-Kunt, A. (2021). Small and medium-size enterprises: Access to finance as a growth constraint. *Journal of Banking & Finance*, 30(11), 2931-2943. https://doi.org/10.1016/j.jbankfin.2021.106514

Biblioteca de la Universidad de Navarra. (2023, 6 marzo). ¿Qué es una revisión sistemática? ¿Cómo realizarla? Pasos o Etapas [Vídeo]. YouTube. https://www.youtube.com/watch?v=5_gY5HypOLg

Bishop, C. M. (2023). *Deep Learning: Foundations and Concepts. Springer*.

Brenning, A. (2021). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package 'sperrorest'. *Environmental Modelling & Software*, 145, 105179.

Casali, Y., Aydin, N. Y., & Comes, T. (2022). Machine learning for spatial analyses in urban areas: A scoping review. *Sustainable Cities and Society*, 85, 104050. https://doi.org/10.1016/j.scs.2022.104050

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

Comisión Económica para América Latina y el Caribe. (2024). *Índice Latinoamericano de Inteligencia Artificial (ILIA)*. https://www.cepal.org/es/publicaciones/indice-latinoamericano-inteligencia-artificial

Confecámaras. (2023). *La supervivencia empresarial en Colombia: estudio de los factores clave que impulsan la permanencia de las empresas en el mercado*. https://confecamaras.org.co/estudio-supervivencia-empresarial/

Confecámaras. (2023). *Reporte de Empresas y Empleo*. Confederación Colombiana de Cámaras de Comercio. https://www.confecamaras.org.co

DANE. (2024). *Caracterización del ecosistema de datos en Colombia*. https://infraestructuradatos.gov.co/798/articles-383921_recurso_1.pdf

de Lema, G. P., & Segura, C.-F. (2021). *Impacto económico de la CRISIS COVID-19 sobre la mipyme en Iberoamérica*.

Departamento Nacional de Planeación (DNP). (2021). *Documento Conpes 4026: Política Nacional para la Gestión del Suelo Urbano*.

Díaz, L. I., & Ramírez, T. D. (2021). *El emprendimiento, la transformación digital e innovación en Colombia como oportunidad de desarrollo económico*.

Díaz, Z., Fernández, J., & Segovia, M. J. (2004). Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras. *Documentos de Trabajo, Universidad Complutense de Madrid*. https://doi.org/10.5209/rev_PRE.2004.v8.art9

El Meridiano. (2024). Mortalidad empresarial en Colombia: El 24% de las empresas no sobrevive el primer año. https://elmeridiano.co/cordoba/opinion/mortalidad-empresarial-en-colombia

Godaddy. (2024). Las pymes en Colombia: Claves tecnológicas para crecer en 2024. https://www.godaddy.com/resources/latam/emprender/observatorio-digitalizacion-2024-tecnologia-seguridad-colombia

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning. MIT Press*.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Infraestructura de Datos Espaciales de Colombia. (2024). *IDE Colombia - Capas geográficas disponibles*. https://www.ideca.gov.co

Instituto Geográfico Agustín Codazzi (IGAC). (2023). *Especificaciones para la Gestión de Información Geográfica*.

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

Kitchenham, B., & Charters, S. (2021). *Guidelines for performing Systematic Literature Reviews in Software Engineering* (Version 2.0). EBSE Technical Report.

Klein, R., Popp, B., & Schmidt, S. (2023). Predicting retail store performance using location data and machine learning. *Journal of Retailing and Consumer Services*, 72, 103265. https://doi.org/10.1016/j.jretconser.2023.103265

La Nota Económica. (2024). En Colombia el 91,8% de las empresas son PyMEs y generan 3 millones de empleos. https://lanotaeconomica.com.co/movidas-empresarial/en-colombia-el-918-de-las-empresas-son-pymes/

Li, X., Song, Y., & Liu, Y. (2023). The impact of business agglomeration on new venture survival: A machine learning approach. *Small Business Economics*, 61(4), 1451–1472. https://doi.org/10.1007/s11187-022-00740-4

Lundberg, S. M., & Lee, S. I. (2021). A unified approach to interpreting model predictions. *Nature Machine Intelligence*, 3(10), 1–10.

Matuszelański, K., & Kopczewska, K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(1), Article 1. https://doi.org/10.3390/jtaer17010009

MDc. (2025). *Estado del Tejido Empresarial en Colombia 2025: Densidad por departamento*. https://mdc.org.co/tejido-empresarial-colombia-2025/

MDPI. (2025). Predicting business failure with the XGBoost algorithm. *Sustainability*, 17(11), 4948. https://doi.org/10.3390/su17114948

Ministerio de Tecnologías de la Información y las Comunicaciones. (2023). *Política Nacional de Datos*. https://www.mintic.gov.co

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Chapman and Hall/CRC.

Molitor, D., Gupta, S., & Riquelme, C. (2023). Machine learning for small business success prediction: A review. *Journal of Business Analytics*, 6(2), 45–62. https://doi.org/10.1080/2573234X.2023.2187562

Moya-Gómez, B., & García-Palomares, J. C. (2021). The accessibility of retail activities in the city centre: A GIS-based analysis for a more inclusive urban planning. *Journal of Urban Management*, 10(4), 365–378. https://doi.org/10.1016/j.jum.2021.08.001

Roman, D. (2021). *Supervivencia de las nuevas empresas*.

Romero, M., Carmona, P., & Pozuelo, J. (2024). Ethical transparency in business failure prediction: Uncovering the black box of XGBoost algorithm. *Spanish Journal of Finance and Accounting/Revista Española de Financiación y Contabilidad*., 54, 135-165. https://doi.org/10.1080/02102412.2024.2301234

Stam, E. (2021). *Handbook of Research on Business and Technology Impacts on Knowledge Management.*

U.S. National Library of Medicine. (2020). *Medical Subject Headings* (MeSH). National Institutes of Health. https://www.nlm.nih.gov/mesh/meshhome.html

Vera, S. A. A., Segovia, M. M. Z., Cevillano, A. N. Z., & Ponce, V. L. M. (2024). Técnicas de machine learning aplicadas a la interpretación de datos de mercado. *Ciencia y Desarrollo*, 27(2), Article 2. https://doi.org/10.21503/cyd.v27i2.2615

Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, 58(4), 102555. https://doi.org/10.1016/j.ipm.2021.102555

Zhou, Z.-H. (2021). *Ensemble Methods: Foundations and Algorithms* (2nd ed.). Chapman and Hall/CRC.