Universidad Continental

Escuela de Posgrado

MAESTRÍA EN CIENCIA DE DATOS

Tesis

Predicción de la insolvencia de cajas municipales y rurales en el Perú mediante técnicas de machine learning

Angela Adriana Rivera Delgado

Para optar el Grado Académico de
Maestro en Ciencia de Datos

Lima, 2025

Universidad Continental

ANEXO 6

INFORME DE CONFORMIDAD DE ORIGINALIDAD DEL

TRABAJO DE INVESTIGACIÓN

A	:	Mg. Jaime Sobrados Tapia Director Académico de la Escuela de Posgrado
DE	:	**Kevin Rafael Palomino Pacheco** Asesor del Trabajo de Investigación
ASUNTO	:	Remito resultado de evaluación de originalidad de Trabajo de Investigación
FECHA	:	8 de agosto del 2025

Con sumo agrado me dirijo a vuestro despacho para saludarlo y en vista de haber sido designado Asesor del Trabajo de Investigación/Tesis/Artículo Científico titulado "**Predicción de la insolvencia de cajas municipales y rurales en el Perú mediante técnicas de machine learning**", perteneciente a **Bach. Angela Adriana Rivera Delgado**, de la **MAESTRÍA EN CIENCIA DE DATOS**; se procedió con el análisis del documento mediante la herramienta "Turnitin" y se realizó la verificación completa de las coincidencias resaltadas por el software, cuyo resultado es **11%** de similitud (informe adjunto) sin encontrarse hallazgos relacionados con plagio. Se utilizaron los siguientes filtros:

- Filtro de exclusión de bibliografía SÍ [x] NO [ ]
- Filtro de exclusión de grupos de palabras menores SÍ [x] NO [ ]

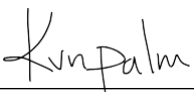
  (Máximo nº de palabras excluidas: < **20**)
- Exclusión de fuente por trabajo anterior del mismo estudiante SÍ [x] NO [ ]

En consecuencia, se determina que el trabajo de investigación constituye un documento original al presentar similitud de otros autores (citas) por debajo del porcentaje establecido por la Universidad.

Recae toda responsabilidad del contenido de la tesis sobre el autor y asesor, en concordancia a los principios de legalidad, presunción de veracidad y simplicidad, expresados en el Reglamento del Registro Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales – RENATI y en la Directiva 003-2016-R/UC.

Esperando la atención a la presente, me despido sin otro particular y sea propicia la ocasión para renovar las muestras de mi especial consideración.

Atentamente,



Kevin Rafael Palomino Pacheco
DNI: 1045711819

Universidad Continental

DECLARACIÓN JURADA DE AUTENTICIDAD

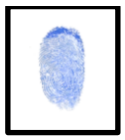
Yo, ANGELA ADRIANA RIVERA DELGADO, identificada con Documento Nacional de Identidad N° 70672028, egresada de la MAESTRÍA EN CIENCIA DE DATOS, de la Escuela de Posgrado de la Universidad Continental, declaro bajo juramento lo siguiente:

1. La Tesis titulada "PREDICCIÓN DE LA INSOLVENCIA DE CAJAS MUNICIPALES Y RURALES EN EL PERÚ MEDIANTE TÉCNICAS DE MACHINE LEARNING", es de mi autoría, el mismo que presento para optar el Grado Académico de MAESTRO EN CIENCIA DE DATOS.
2. La Tesis no ha sido plagiada ni total ni parcialmente, para lo cual se han respetado las normas internacionales de citas y referencias para las fuentes consultadas, por lo que no atenta contra derechos de terceros.
3. La Tesis es original e inédita, y no ha sido realizado, desarrollado o publicado, parcial ni totalmente, por terceras personas naturales o jurídicas. No incurre en autoplagio; es decir, no fue publicado ni presentado de manera previa para conseguir algún grado académico o título profesional.
4. Los datos presentados en los resultados son reales, pues no son falsos, duplicados, ni copiados, por consiguiente, constituyen un aporte significativo para la realidad estudiada.

De identificarse fraude, falsificación de datos, plagio, información sin cita de autores, uso ilegal de información ajena, asumo las consecuencias y sanciones que de mi acción se deriven, sometiéndome a las acciones legales pertinentes.

Arequipa, 12 de mayo de 2025

ANGELA ADRIANA RIVERA DELGADO
DNI. N° 70672028



Huella

PREDICCIÓN DE LA INSOLVENCIA DE CAJAS MUNICIPALES Y RURALES EN EL PERÚ MEDIANTE TÉCNICAS DE MACHINE LEARNING

INFORME DE ORIGINALIDAD

11%	9%	5%	6%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Submitted to Universidad Continental Trabajo del estudiante	2%
2	repositorio.continental.edu.pe Fuente de Internet	1%
3	hdl.handle.net Fuente de Internet	1%
4	export.arxiv.org Fuente de Internet	1%
5	revistas.uandina.edu.pe Fuente de Internet	<1%
6	www.coursehero.com Fuente de Internet	<1%
7	www.ibm.com Fuente de Internet	<1%
8	intellectum.unisabana.edu.co Fuente de Internet	<1%

9	Submitted to TecnoCampus Trabajo del estudiante	<1 %
10	Submitted to Universidad Internacional de la Rioja Trabajo del estudiante	<1 %
11	eprints.ucm.es Fuente de Internet	<1 %
12	Submitted to udep Trabajo del estudiante	<1 %
13	repositorio.unajma.edu.pe Fuente de Internet	<1 %
14	www.clubensayos.com Fuente de Internet	<1 %
15	Submitted to Universidad del Valle de Guatemala Trabajo del estudiante	<1 %
16	dspace.utb.edu.ec Fuente de Internet	<1 %
17	Salamanca Vásquez, Esteban. "Identificación de Patrones de Deserción y Riesgo Académico en Carreras de la Facultad de Ingeniería de la Universidad Distrital a Través de Técnicas Machine Learning", Universidad Distrital Francisco José de Caldas (Colombia) Publicación	<1 %

18	repositorio.ucv.edu.pe Fuente de Internet	<1 %
19	Vicente Marlon Villa Villa, Rodrigo Enrique Velarde Flores, Mayra Karina Flores Escobar, Jhonny Javier León Cando. "Assessing the solvency of footwear factories in Tungurahua: an analysis based on the Z-Score model", Esprint Investigación, 2024 Publicación	<1 %
20	Submitted to consultoriadeserviciosformativos Trabajo del estudiante	<1 %
21	Submitted to UNIBA Trabajo del estudiante	<1 %
22	repositorio.uchile.cl Fuente de Internet	<1 %
23	dspace.unl.edu.ec Fuente de Internet	<1 %
24	Submitted to Universidad Autónoma de Nuevo León Trabajo del estudiante	<1 %
25	core.ac.uk Fuente de Internet	<1 %
26	rodin.uca.es Fuente de Internet	<1 %

27	Submitted to Universidad Tecnica De Ambato-Direccion de Investigacion y Desarrollo , DIDE Trabajo del estudiante	<1 %
28	ebuah.uah.es Fuente de Internet	<1 %
29	Submitted to uncedu Trabajo del estudiante	<1 %
30	Caceres Bustinza, Lycet Maria. "La inversión y su incidencia en la rentabilidad de las cajas municipales de ahorro y crédito del sur del Perú en la década 2007-2016", Universidad Nacional del Altiplano de Puno (Peru) Publicación	<1 %
31	Submitted to Universidad TecMilenio Trabajo del estudiante	<1 %
32	repositorio.unitec.edu Fuente de Internet	<1 %
33	Quintana Quispe, José Orlando. "Análisis predictivo de la deserción estudiantil en los alumnos de la Universidad Nacional de Moquegua entre 2009 y 2019, usando Machine Learning", Universidad Nacional del Altiplano de Puno (Peru) Publicación	<1 %
34	Submitted to Universidad Nacional Abierta y a Distancia, UNAD,UNAD	<1 %

Trabajo del estudiante

35	riunet.upv.es Fuente de Internet	<1 %
36	Submitted to Universidad Carlos III de Madrid - EUR Trabajo del estudiante	<1 %
37	repositorioacademico.upc.edu.pe Fuente de Internet	<1 %
38	tarija200.com Fuente de Internet	<1 %
39	Chavesta Ramos, Kelly | De Rutte Torres, Eduardo | Valencia Castillo, Victor | Zavala Sosa, Marco. "Planeamiento Estrategico del Sistema de Cajas Municipales del Peru", Pontificia Universidad Catolica del Peru - CENTRUM Catolica (Peru), 2021 Publicación	<1 %
40	Submitted to antonionarino Trabajo del estudiante	<1 %

Excluir citas	Apagado	Excluir coincidencias	< 20 words
Excluir bibliografía	Activo		

Asesor

Dr. Kevin Rafael Palomino Pacheco

Agradecimientos

Deseo expresar mi más profundo y sincero agradecimiento a todas las personas que hicieron posible la realización de este trabajo. En especial, a mis padres, quienes han sido mi mayor fortaleza y guía en cada etapa de mi vida. Gracias por creer en mí incluso cuando yo dudaba, por su amor incondicional y por los innumerables sacrificios que han hecho para que yo pudiera alcanzar mis metas.

Índice

Índice de Tablas

Índice de Figuras

Resumen

El sistema financiero peruano enfrenta retos en la sostenibilidad de las cajas municipales y rurales, que son fundamentales para la inclusión financiera de las zonas rurales y semiurbanas. Esta investigación identifica un modelo predictivo utilizando técnicas de Machine Learning para anticipar la insolvencia en estas instituciones, lo que aborda un problema clave en el fortalecimiento del sector micro financiero. Se identificaron variables críticas para predecir la insolvencia, como rentabilidad, eficiencia, calidad de activos, liquidez e inflación. Tras evaluar y comparar varios métodos de Machine Learning, el modelo Random Forest resultó ser el más robusto y eficiente, alcanzando una precisión del 75% y un F1-Score Macro promedio de 0.75. Este modelo superó a otros enfoques tradicionales, como la regresión logística y a algoritmos como Árboles de Decisión y SVM (Support Vector Machine), destacándose por su capacidad para manejar relaciones no lineales y datos desbalanceados. Los hallazgos subrayan la relevancia de integrar herramientas avanzadas en la gestión de riesgos financieros, ofreciendo soluciones prácticas para mejorar la sostenibilidad y la confianza en el sistema micro financiero peruano.

Palabras clave: Machine Learning, cajas municipales, cajas rurales, solvencia financiera, modelo predictivo.

Abstract

The Peruvian financial system faces challenges in the sustainability of municipal and rural savings banks, which are crucial for financial inclusion in rural and semi-urban areas. This research identifies a predictive model using Machine Learning techniques to anticipate insolvency in these institutions, addressing a key issue in strengthening the microfinance sector. Critical variables for predicting insolvency were identified, including profitability, efficiency, asset quality, liquidity, and inflation. After evaluating and comparing several Machine Learning methods, the Random Forest model was found to be the most robust and efficient, achieving 75% accuracy and an average F1-Score Macro of 0.75. This model outperformed other traditional approaches, such as logistic regression, as well as algorithms like Decision Trees and SVM (Support Vector Machine), excelling in handling non-linear relationships and imbalanced data. The findings highlight the importance of integrating advanced tools in financial risk management, providing practical solutions to enhance sustainability and trust in the Peruvian microfinance system.

Keywords: Machine Learning, municipal savings banks, rural savings banks, financial solvency, predictive model.

Introducción

En el Perú, las cajas municipales y rurales juegan un papel vital en la inclusión financiera y en el desarrollo de las comunidades más vulnerables. Estas instituciones acercan los servicios financieros a zonas rurales y semiurbanas, donde muchas personas y pequeños negocios o microempresas carecen de acceso a los recursos administrados por la banca tradicional. Más allá de facilitar créditos y ahorros, las cajas municipales y rurales son pilares fundamentales para dinamizar las economías locales, fomentar el emprendimiento y mejorar la calidad de vida de miles de peruanos. Sin embargo, detrás de esta labor esencial, enfrentan riesgos importantes, entre ellos, la posibilidad de insolvencia, un problema que, si no se gestiona adecuadamente, puede afectar tanto a las instituciones como a los usuarios que dependen de ellas.

La predicción de insolvencia, por tanto, no solo es una herramienta técnica, sino también una estrategia crucial para garantizar la sostenibilidad del sistema financiero. Poder anticipar riesgos permite a las instituciones actuar de manera preventiva, optimizar recursos y tomar decisiones estratégicas más acertadas. En este escenario, las técnicas de Machine Learning han demostrado ser herramientas poderosas y revolucionarias. A diferencia de los métodos tradicionales, que a menudo se limitan a suposiciones lineales, el Machine Learning permite descubrir patrones complejos y manejar grandes volúmenes de datos, lo que lo hace una herramienta óptima para examinar las complejidades del ámbito financiero.

A escala global, investigaciones recientes han evidenciado la ventaja de los modelos de Machine Learning en la predicción de insolvencias financieras, destacando algoritmos como Random Forest, XGBoost y redes neuronales en países como Estados Unidos, Colombia, Argentina y Ecuador, donde se han combinado variables financieras y no financieras con éxito. En cuanto a Perú, las investigaciones sobre solvencia en instituciones microfinancieras han adoptado enfoques tradicionales, como el modelo Z-Score de Altman, dejando una brecha con relación al uso de técnicas modernas.

La presente investigación tiene como objetivo identificar un modelo predictivo construido en base a técnicas de Machine Learning que permita anticipar de manera precisa la insolvencia en las cajas municipales y rurales del sistema financiero del Perú. Para lograrlo, se analizó las variables macroeconómicas y microfinancieras más relevantes, se evaluaron los métodos de Machine Learning más efectivos y se compararon sus resultados con los obtenidos a partir de enfoques tradicionales. A través de este modelo, se proporcionar una herramienta práctica y efectiva que no solo fortalece la gestión de riesgos en estas instituciones, sino que también contribuye al desarrollo sostenible del sector micro financiero peruano.

La estructura de este trabajo de investigación se presenta de la siguiente manera: en el primer capítulo se expone el planteamiento del problema a tratar, los objetivos y la justificación del estudio realizado; en el segundo capítulo se muestra la revisión de los antecedentes investigativos y el marco teórico, destacando la evolución de la aplicación de las técnicas de Machine Learning en el análisis propio del sector financiero. El tercer capítulo presenta las hipótesis y variables como punto inicial de la investigación; como complemento el cuarto capítulo desarrolla la parte metodológica y su operacionalización. En el quinto capítulo se desarrolla la analítica de los datos, desde la selección de estos, su estandarización y la obtención de los diferentes modelos que permite arribar a los resultados y la comparación del rendimiento de los diferentes métodos evaluados. Finalmente, el sexto capítulo resume las conclusiones principales y formula recomendaciones prácticas para implementar los hallazgos de esta investigación.

Este estudio busca ofrecer soluciones concretas para un problema crítico en el sistema micro financiero peruano. Al adoptar modelos predictivos avanzados, las cajas municipales y rurales podrán no solo reducir riesgos, sino también garantizar su sostenibilidad a largo plazo. Asimismo, sienta las bases para futuros estudios que busquen integrar técnicas más avanzadas y explorar variables adicionales en el análisis de riesgos financieros en el Perú.

Capítulo I: Planteamiento del estudio

1.1. Planteamiento y formulación del problema

1.1.1. Planteamiento del problema.

A nivel global, las microfinanzas han sido fundamentales para impulsar la inclusión financiera, especialmente en economías emergentes donde el acceso a la banca tradicional sigue siendo limitado. De acuerdo con los datos del Global Findex, 1.4 mil millones de personas en todo el mundo aún no obtienen servicios bancarios formales, entonces es evidente la importancia de las instituciones microfinancieras y su accionar para disminuir la exclusión financiera, fomentar el ahorro, y brindar apoyo crediticio a emprendedores y pequeñas empresas en zonas vulnerables (WORLD BANK GROUP, 2021).

En América Latina, la exclusión financiera sigue siendo un problema significativo. Según el Banco Interamericano de Desarrollo, en 2022, el 45% de los adultos de la región carecían de una cuenta bancaria formal, por lo que sigue siendo necesario desarrollar y potenciar las instituciones microfinancieras (Andrade, Herrera, & De Olloqui, 2015). Sin embargo, el rol de estas instituciones se ve amenazado por los posibles problemas económicos de los clientes, que se traduce en morosidad elevada que podría devenir en la quiebra o insolvencia de las mismas. Así, por ejemplo, en México, en 2023, la tasa de morosidad en las instituciones microfinancieras alcanzó el 8%, siendo la primera vez que superó el 7,5% desde 2017, lo que obligó a las autoridades a endurecer las políticas de crédito (Alvarez, 2023). En Colombia, la tasa de morosidad en las instituciones microfinancieras aumentó del 5,6% al 6,3% entre 2022 y 2023, mostrando la vulnerabilidad de este sector ante fluctuaciones económicas (SFC, 2021).

En el Perú, las cajas municipales son instituciones fundamentales para incentivar la inclusión financiera, en particular en áreas rurales y ámbitos regionales, ya que la banca tradicional es menos común. Además de promover el ahorro, las cajas también prestan crédito a las micro, pequeñas y medianas empresas (MIPYME)

que conforman el 99.4% de la estructura empresarial del país y generan el 61% del empleo formal, siendo fundamentales para el desarrollo económico local (PRODUCE, 2024).

En los últimos años ha sido considerable el crecimiento de las MIPYME en el Perú, por ejemplo, en el 2022, el número de MIPYME ascendió a más de 2.2 millones, lo que representó un incremento del 6% con respecto al año anterior. Además, estas empresas contribuyeron con aproximadamente S/337 000 millones en ventas durante ese mismo año (PRODUCE, 2024). Este crecimiento se ha visto impulsado por las políticas de inclusión financiera, especialmente a través de las cajas municipales y rurales que brindan acceso al crédito a miles de emprendedores en todo el país.

Sin embargo, en los últimos meses, las cajas municipales enfrentaron una fuerte crisis, con una reducción drástica en sus utilidades; en abril del 2024, las utilidades anuales de las cajas municipales se desplomaron en un 77.8% en comparación con el mismo período del año anterior, pasando de S/65.2 millones en 2023 a solo S/14.5 millones en el 2024 (GanaMas, 2024).

Algunas de las cajas más afectadas por esta situación fueron Caja Sullana, que reportó pérdidas netas de S/33.7 millones y Caja Tacna, con una pérdida de S/12.4 millones (GanaMas, 2024). Estas cifras reflejan la vulnerabilidad de las cajas frente a fluctuaciones económicas, especialmente en un contexto de creciente inflación y aumento de tasas de interés, que afectan directamente la capacidad de pago de sus clientes.

El índice de morosidad en las cajas municipales también ha mostrado un incremento preocupante. En enero de 2024, la morosidad global de estas entidades alcanzó el 6.1%, con algunas cajas como Caja Sullana registrando niveles alarmantes de 17.7%, Caja Tacna con 9.6%, y Caja del Santa con 8.3% (GanaMas, 2024). Estos altos niveles de morosidad sugieren una inadecuada gestión de riesgos y una creciente exposición al sobreendeudamiento de sus clientes.

Las cajas rurales, por otro lado, presentan una situación aún más delicada debido a su menor tamaño y capacidad operativa. Estas instituciones, que también juegan un rol fundamental en la inclusión financiera de las zonas rurales, han mostrado un aumento en los índices de morosidad, especialmente en los sectores agrícolas y microempresariales. Además, la menor diversificación de su cartera de clientes, combinada con su alta dependencia de sectores más vulnerables a la volatilidad económica, como en el caso de la agricultura, las hacen especialmente susceptibles a las fluctuaciones externas

Factores como el endeudamiento excesivo de sus clientes y una gestión deficiente del riesgo crediticio, junto con el impacto de la fluctuación macroeconómica, han llevado a algunas de estas instituciones a ser incapaces de cumplir con sus obligaciones financieras. Esto no solo afecta su propio potencial de sostenibilidad, sino también la confianza del público en general en el sistema financiero descentralizado. La insolvencia de las cajas puede, de hecho, tener un impacto en cascada a través de la economía: desde la pérdida de los ahorros de los depositantes hasta el colapso de una fuente crucial de crédito para miles de pequeños emprendedores.

Actualmente, se pronostica la solvencia de estas instituciones financieras en base a evaluaciones históricas e indicadores netamente financieros, sin embargo estos métodos tradicionales no siempre detectan a tiempo las desviaciones y problemas, más aún en el contexto peruano, donde las fluctuaciones macroeconómicas, como el aumento de los precios o cambios en las tasas de interés, tienen un gran impacto en la capacidad de pago de los clientes de las cajas y por ende en la capacidad de pago de la cajas.

Ante este contexto, el uso de la ciencia de datos, específicamente de los métodos de Machine Learning, surge como una alternativa innovadora para desarrollar modelos predictivos de la insolvencia de estas instituciones financieras. Estos métodos se basan en analizar grandes volúmenes de datos, para así identificar patrones y relaciones, y generar predicciones más precisas. A diferencia de los métodos tradicionales, los métodos de Machine Learning mejoran continuamente

con la entrada de nuevos datos, lo que permite identificar los factores de riesgo en fases más tempranas, antes de que se conviertan en problemas críticos. Algoritmos como la regresión logística, árboles de decisión, modelos de boosting y redes neuronales son herramientas clave para poder desarrollar estos modelos predictivos.

Por todo lo mencionado, el principal problema es la falta de un modelo predictivo que utilice métodos de Machine Learning para prever la insolvencia de las cajas municipales y rurales del Perú, que utilice los datos micro financieros y macroeconómicos disponibles, ofreciendo una visión más completa de la salud financiera de estas instituciones.

1.1.2. Formulación del problema.

A. Problema general.

¿Se puede identificar un modelo predictivo que permita anticipar la insolvencia en las cajas municipales y rurales del Perú utilizando métodos de Machine Learning?

B. Problemas específicos

- ¿Cuáles son las variables macroeconómicas y microfinancieras más relevantes para predecir la insolvencia en cajas municipales y rurales del Perú?

- ¿Cuál es el método de Machine Learning más efectivo para predecir la insolvencia en las cajas municipales y rurales del Perú?

- ¿Qué diferencias existen en el rendimiento predictivo de los métodos de Machine Learning con los métodos tradicionales para predecir la insolvencia en las cajas municipales y rurales del Perú?

1.2. Determinación de objetivos

1.2.1. Objetivo general.

Identificar un modelo predictivo que permita anticipar la insolvencia en las cajas municipales y rurales del Perú utilizando métodos de Machine Learning.

1.2.2. Objetivos específicos.

- Determinar las variables macroeconómicas y microfinancieras más relevantes para predecir la insolvencia en las cajas municipales y rurales del Perú, utilizando métodos de Machine Learning.

- Evaluar y seleccionar el método de Machine Learning más efectivo para predecir la insolvencia en las cajas municipales y rurales del Perú.

- Comparar el rendimiento predictivo de los métodos de Machine Learning con los enfoques tradicionales de predicción de insolvencia en las cajas municipales y rurales del Perú.

1.3. Justificación e importancia del estudio

1.3.1. Justificación teórica.

Este estudio pretende profundizar en el mundo de los modelos de predicción de insolvencia en las microfinanzas, utilizando técnicas avanzadas de Machine Learning que no han sido ampliamente exploradas en el contexto de las cajas municipales y rurales. Si bien la mayoría de los bancos utilizan modelos predictivos para evaluar su estabilidad financiera, las instituciones de microfinanzas, que trabajan en entornos desafiantes, no han tenido muchos estudios que utilicen métodos avanzados para analizar su solvencia. Este trabajo tiene como objetivo abordar las limitaciones de los métodos tradicionales mediante el uso de algoritmos de Machine Learning, logrando predicciones más precisas al tener en cuenta no

solo factores micro financieros, sino también factores macroeconómicos. Esta contribución teórica ofrece una visión más amplia, adaptada a los desafíos únicos que enfrentan las economías emergentes, que ayudará a mejorar la estabilidad financiera y la sostenibilidad en las microfinanzas, al tiempo que garantiza la inclusión financiera de las zonas descentralizadas.

1.3.2. Justificación metodológica.

Desde una perspectiva metodológica, este estudio está respaldado en el uso de métodos avanzados de Machine Learning para evaluar la probabilidad de insolvencia de las cajas municipales y rurales del Perú. Mediante el uso de modelos como árboles de decisión y redes neuronales, se tiene como objetivo realizar predicciones más precisas que las obtenidas por métodos tradicionales, utilizando mayor cantidad de datos y variedad de variables independientes, como son los datos macroeconómicos. Esta metodología permite tener herramientas más sólidas de predicción de insolvencia para instituciones financieras, las cuales podrán ser replicadas y /o adaptadas para futuras investigaciones en diferentes situaciones problemáticas

1.3.3. Justificación social.

La justificación social del estudio radica en la capacidad de aumentar la estabilidad financiera de las cajas municipales y rurales, que son fundamentales para la inclusión financiera de zonas descentralizadas del País, donde la banca tradicional no llega. Estas instituciones desempeñan un papel importante al proporcionar créditos a los ciudadanos más vulnerables y sobre todo brindar oportunidades de crecimiento a los micro y pequeños emprendedores. Al poder aplicar modelos predictivos de Machine Learning para anticipar la insolvencia de las cajas permite mejorar la seguridad y salud financiera y, en consecuencia, permite la sostenibilidad a largo plazo de estas. Además, los resultados podrían respaldar políticas públicas enfocadas en el desarrollo económico de zonas más vulnerables y con poco acceso a la banca tradicional.

1.4. Limitaciones de la presente investigación

La presente investigación presenta dos limitaciones. En primer lugar, se tiene la calidad de los datos históricos que pueden impactar en la precisión del modelo predictivo, debido a que estos datos pueden estar incompletos o ser inconsistentes. El segundo limitante son los cambios en el entorno macroeconómico que pueden alterar las variables macroeconómicas que afectan la solvencia, como variaciones en el ámbito político, el ámbito legal o crisis externas a las instituciones, afectando la aplicabilidad del modelo a largo tiempo.

Capítulo II: Marco teórico

2.1. Antecedentes de la investigación

2.1.1. Internacionales.

La predicción de la solvencia de diversas instituciones mediante técnicas de Machine Learning ha surgido como un área de investigación fundamental en el contexto financiero global, evidenciando la necesidad de poder anticipar crisis financieras para poder reaccionar y tomar decisiones que permitan la sostenibilidad de las instituciones en este rubro. En este contexto, algunos estudios recientes revelan métodos innovadores y eficientes para poder realizar predicciones financieras, mientras que otros estudios enriquecen la comprensión de las variables que influyen en la solvencia de las instituciones.

En el año 2024, un artículo realizado por un grupo de investigadores en Argentina mostró un modelo predictivo que buscaba anticipar la insolvencia empresarial utilizando y comparando técnicas de análisis discriminante lineal y árboles de clasificación, en base a datos obtenidos del Mercado de Valores de Buenos Aires. Como resultado de la investigación, se concluyó que los métodos no lineales tienen mayor precisión en la predicción que los métodos tradicionales. Además, se evidenció que una variable fundamental para la evaluación de la solvencia es el ratio de la capacidad de pago en diferentes momentos. Estos resultados son valiosos para la presente investigación, debido a que enfatiza la importancia de utilizar técnicas más avanzadas de predicción, que se adapten mejor al contexto financiero y económico peruano, donde las dinámicas de insolvencia empresarial pueden diferir significativamente de las observadas en el contexto argentino (Sattler, Terreno, & Pérez, 2024).

Este mismo año, Yuly Franco publicó un artículo en Colombia, donde se desarrolla una revisión sistemática de 300 documentos relacionados con la predicción de quiebras empresariales usando Machine Learning. Analizando los resultados obtenidos con algoritmos como XGBoost, SVM y Random Forest, se encontró que

estos métodos no solo superan a las técnicas tradicionales, sino que también muestran una mayor eficacia cuando la predicción se realiza con un año de anticipación. En este artículo, también se evidencia que, para una mejor predicción de quiebra empresarial, es fundamental usar tanto variables financieras como variables no financieras. Este hallazgo es particularmente relevante para el desarrollo de un modelo predictivo para las cajas municipales y rurales en Perú, debido a que denota la importancia de un enfoque integral que considere diversas variables para la evaluación del riesgo (Franco, 2024).

Por otro lado, en el año 2023, Alison Jara realizó un estudio en Ecuador con el propósito de analizar la posibilidad de quiebra en cooperativas de ahorro y crédito, utilizando el modelo Z-Score de Altman. La investigación, de enfoque netamente cuantitativo, examinó indicadores financieros clave como liquidez, rentabilidad y eficiencia. Los resultados indicaron que una proporción considerable de estas cooperativas presenta dificultades financieras, evidenciadas por valores Z menores a 2, lo que sugiere un alto riesgo de insolvencia. En este sentido, dichos hallazgos son relevantes para el presente estudio, ya que proporcionan una referencia analítica para evaluar la estabilidad financiera de instituciones similares en Perú, destacando que el uso de modelos de Machine Learning podría optimizar la identificación de factores de riesgo en el sector financiero peruano. (Jara Velastegui, 2023).

En el año 2022, un artículo presentado por un grupo de investigadores en Italia muestra cómo se examinaron mediante diversas técnicas de Machine Learning a los bancos de la región dentro de un periodo de dos años (del 2018 al 2020) con la finalidad de determinar los factores del impago bancario. En base a esto, se obtuvo un efectivo modelo predictivo con redes neuronales, el cual revela la importancia de los métricas financieras en la evaluación de riesgos de impago bancario. Este enfoque metodológico es significativo para el estudio en curso, ya que permite la adaptación de técnicas avanzadas en el análisis de la insolvencia de las cajas municipales y rurales en el Perú (Valentina Lagasio, 2022).

En 2020, Anastasios Petropoulos publicó un artículo en Estado Unidos sobre insolvencia bancaria, donde muestra los resultados de aplicar varios modelos de predicción, como la regresión logística, el análisis discriminante lineal, Random Forests y redes neuronales, empleando datos de la Federal Deposit Insurance Corporation (FDIC). Los resultados mostraron que el modelo predictivo de insolvencia con mejor rendimiento es el de Random Forests. Además, se identificaron como variables clave a los ingresos y el capital dentro del marco CAMELS. Si bien este artículo se centra en bancos estadounidenses, las metodologías son ajustables en el contexto peruano, donde la aplicación de técnicas similares podría ofrecer valiosas perspectivas sobre el riesgo de insolvencia en las cajas municipales y rurales (Anastasios Petropoulos, 2020).

Las investigaciones a nivel internacional proporcionan una base sólida que resalta la importancia de aplicar técnicas de Machine Learning en la predicción de insolvencias dentro del sector financiero. La combinación de enfoques avanzados y el análisis de múltiples variables, tanto financieras como no financieras, resulta fundamental para prever con precisión los riesgos que afectan la estabilidad financiera de las cajas municipales y rurales en el Perú.

2.1.2. Nacionales.

En el contexto peruano, en el año 2024, Rossy Medina presentó en su tesis de grado, un estudio donde buscó identificar los factores determinantes que impactaron en la sostenibilidad de las instituciones durante el período 2000-2019. Para esto utilizó un enfoque cuantitativo y un diseño no experimental, al aplicar modelos de regresión OLS y análisis de panel, analizando variables cruciales como la inflación, el ingreso per cápita, la eficiencia operativa y la cartera en riesgo. El principal hallazgo muestra que la autosuficiencia operativa, un indicador fundamental de sostenibilidad, se ve afectada por la mayoría de las variables estudiadas, salvo la actividad económica. Asimismo, mostró que, aunque la rentabilidad es esencial, no garantiza la sostenibilidad a largo plazo, lo que sugiere que las microfinancieras deben buscar un equilibrio entre estos aspectos para asegurar su éxito. Este estudio se relaciona directamente con la investigación

actual, ya que ambos resaltan la importancia de la sostenibilidad financiera, aunque el enfoque aquí utilizado va más allá al proponer un modelo predictivo utilizando técnicas de Machine Learning, lo que introduce una metodología más sofisticada en la predicción de posibles problemas de insolvencia (Medina Huaman, 2024).

Otro estudio relevante es el de Nick Valdivia y Jean Rojas, realizado en 2023, que analiza la situación financiera de las empresas del sector financiero del Perú utilizando el modelo Z-Score de Altman. Esta investigación utiliza un enfoque cuantitativo y un diseño comparativo entre los años 2019 al 2021, evaluó la solvencia de 41 empresas, incluidas bancos y cajas municipales. Los resultados indican que, aunque un 46% de las empresas mostraron un retroceso en su desempeño durante 2020, muchas lograron recuperarse en 2021; sin embargo, se destaca que algunas cajas municipales presentaron un mayor riesgo de insolvencia, lo que resalta la necesidad de implementar medidas efectivas de gestión de liquidez. Este estudio es de particular relevancia para la investigación actual, pues ambos buscan medir la solvencia financiera. Asimismo, la presente investigación al incorporar técnicas de Machine Learning, permite integrar un mayor número de variables y realizar predicciones más precisas sobre la insolvencia, marcando un avance significativo en el análisis del riesgo financiero (Valdivia Saavedra & Rojas Munares, 2023).

Como tercer antecedente, se tiene el artículo sobre el análisis del riesgo de quiebra de instituciones financieras peruanas, presentando por Ricardo Rossi en 2022, en el que también se obtienen aportes al entendimiento de la sostenibilidad financiera. Este estudio utiliza el modelo Z de Altman con la media armónica para evaluar 26 instituciones financieras, encontrando que solo el 20% de las entidades de banca múltiple se encontraba fuera de peligro, mientras que el 13% estaba en riesgo de quiebra. A pesar de que todas las cajas municipales examinadas mostraron un mejor desempeño financiero, sin riesgo de quiebra, la investigación enfatiza la importancia de evaluar los factores de riesgo en este sector. Al igual que en las investigaciones previas, la metodología aplicada en este trabajo es estática, centrada en datos históricos. En contraste, la investigación actual busca desarrollar un enfoque más dinámico utilizando Machine Learning, lo que promete una mayor

precisión en la predicción de insolvencias al analizar grandes volúmenes de datos y considerar tanto variables macroeconómicas como microfinancieras (Rossi Valverde, 2022).

Finalmente, en su tesis de grado del 2023, Haward Chang exploró un enfoque innovador al comparar diversas técnicas de estimación basadas en Machine Learning para la predicción de costos en entidades públicas del Perú. Su estudio aplicó algoritmos de regresión, incluyendo Regresión Lineal Múltiple, Árbol de Decisión, Random Forest y XGBoost, con el objetivo de mejorar la precisión en la estimación de costos dentro de los procesos de contratación pública, demostrando la superioridad de los métodos de Machine Learning sobre las técnicas convencionales. Aunque su investigación se orienta a la gestión pública, el uso de algoritmos avanzados guarda similitudes con la presente propuesta de aplicar Machine Learning para predecir insolvencias en el sector microfinanciero. Mientras Chang se enfoca en la estimación de costos, este estudio extiende la aplicación de modelos como Random Forest y XGBoost para evaluar la estabilidad financiera de las cajas municipales y rurales, con el propósito de desarrollar un modelo predictivo más preciso y relevante. (Chang Hidalgo & Chirinos Mundaca, 2023).

Los estudios analizados subrayan la relevancia de identificar los factores que afectan la sostenibilidad de las cajas municipales y rurales en Perú, evidenciando diversas aproximaciones para su evaluación. La adopción de metodologías más avanzadas, como el Machine Learning, ofrece nuevas posibilidades para optimizar la precisión en la predicción de insolvencias, lo que a su vez facilita el desarrollo de herramientas que contribuyan al fortalecimiento y estabilidad del sector micro financiero en el país.

2.2. Bases teóricas

2.2.1 Desarrollo histórico.

Al inicio de los estudios en el campo de la economía, el análisis de la solvencia financiera se basaba principalmente en métodos tradicionales que utilizaban ratios

financieros estáticos. En este contexto, el modelo Z de Altman o Z-Score, desarrollado en 1968 por el estadounidense Edward Altman, se convirtió en una herramienta fundamental para evaluar la insolvencia empresarial o la posibilidad de quiebre en épocas de crisis. Altman utilizó un enfoque de análisis discriminante, combinando diferentes ratios financieros para predecir la insolvencia, ofreciendo una herramienta valiosa para los analistas financieros (Altman E. , 1968)

Durante los años 1980 y 1990, la investigación se centró en mejorar la precisión de las predicciones de quiebra mediante la aplicación de modelos estadísticos más complejos, al mismo tiempo se comenzó a utilizar la regresión logística como una técnica alternativa. Esta última metodología ofrecía una mayor flexibilidad al permitir la inclusión de variables categóricas y continuas, lo que ampliaba las capacidades analíticas para predecir insolvencias en diferentes contextos económicos (Ohlson, 1980)

Después de 1990, con el avance de la tecnología y la disponibilidad de grandes volúmenes de datos, se comenzó a adoptar metodologías más avanzadas, ya dentro del campo de la inteligencia artificial, como las redes neuronales y los modelos de Machine Learning; metodologías vigentes hoy en día. Estos modelos han demostrado una capacidad superior para manejar grandes volúmenes de datos y patrones complejos, permitiendo una predicción más precisa de la insolvencia financiera. Algoritmos como Support Vector Machines (SVM) y Random Forest han sido adoptados y aplicados en investigaciones recientes, mostrando un rendimiento superior en comparación con los métodos tradicionales. (Umair Ali, 2018)

La crisis financiera del 2008 fue fundamental para la reevaluación de los modelos de riesgo en el sector financiero, esta crisis mostró la vulnerabilidad de muchas instituciones financieras y llevó a un aumento en la investigación sobre la capacidad de predicción de insolvencias. Los estudios comenzaron a integrar no solo variables financieras, sino también factores macroeconómicos y microeconómicos, como el desempleo y la inflación. (Zhou, Keung Lai, & Yen, 2010)

En el contexto peruano, el sistema micro financiero ha enfrentado retos particulares, lo que ha llevado a un aumento en la investigación sobre la salud financiera de las cajas municipales y rurales. Investigaciones recientes han mostrado que, a pesar de la resiliencia del sistema financiero peruano, aún persisten riesgos significativos que pueden poner en peligro la estabilidad de estas instituciones. (Toledo, 2020)

2.2.2 Fundamentación teórica.

La fundamentación teórica de la presente investigación se basa en un conjunto de modelos y metodologías propios del Machine Learning (ML), que se aplican sobre los datos disponibles, con la finalidad de poder predecir la insolvencia empresarial de las cajas rurales y municipalidades en Perú.

Se revisará primero dos modelos estadísticos de predicción de la insolvencia financiera, el modelo Z de Altman y el modelo de Regresión Logística puro, para luego presentar los algoritmos de Machine Learning a ser utilizados

A. Modelo Z de Altman

El profesor Edward Altman, de la Universidad de Nueva York, desarrolló la fórmula Altman Z-Score en la década de 1960. Esta fórmula combina cinco ratios financieros para medir la probabilidad de que una empresa sea insolvente.

$$\text{Altman Z-score} = 1{,}2 * T1 + 1{,}4 * T2 + 3{,}3 * T3 + 0{,}6 * T4 + 1{,}0 * T5 \qquad (1)$$

Donde:

- T1: Capital de trabajo sobre activos totales (WC/TA). Indica la liquidez a corto plazo o liquidez relativa de la empresa
- T2: Utilidades retenidas sobre activos totales (RE/TA). Refleja la capacidad de acumulación de ganancias o su capacidad de reinversión o financiamiento.

- T3: EBIT o utilidades antes de intereses e impuestos sobre activos totales (EBIT/TA). Mide la rentabilidad operativa.
- T4: Valor de mercado del patrimonio o de las acciones sobre pasivos totales (MVE/TL). Evalúa la solvencia a largo plazo.
- T5: Ventas sobre activos totales (S/TA). Indica la eficiencia en el uso de activos, se utiliza como indicador de rotación.

Este modelo ha sido validado y adaptado en diferentes contextos y sectores, incluyendo microfinancieras (Altman E. , 1968). Se emplea como una métrica clave para analizar la insolvencia financiera, tanto en el ámbito académico como en la práctica, permitiendo prever posibles dificultades económicas y el riesgo de insolvencia.

B. Regresión Logística

Es un método de análisis estadístico que emplea modelos matemáticos para establecer relaciones entre dos o más variables, es un análisis de clasificación basado en matemáticas. Esta técnica se utiliza para predecir el valor de una de las variables en función de otras (AWS, 2024). Aunque es un enfoque más tradicional, la regresión logística sigue siendo relevante en el análisis de insolvencia, ya que proporciona interpretaciones claras de los coeficientes y permite evaluar la probabilidad de insolvencia a partir de variables independientes.

Una vez que esta técnica de análisis de datos de clasificación encuentra la relación entre dos o más datos, se utiliza esta relación para predecir el valor de uno de esos factores basándose en los otros, normalmente de forma binaria. Se encuentra dentro de los Modelos Lineales Generalizados (GLM). La importancia de esta técnica radica en su simplicidad, velocidad, flexibilidad y visibilidad.

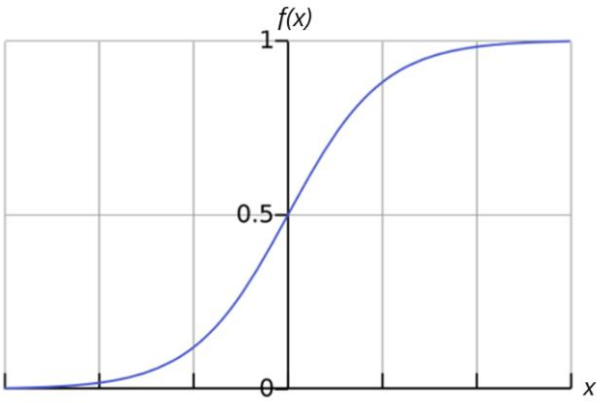
La metodología de aplicación implica definir primero cuáles son las variables que podrían tener relación, recopilar gran cantidad de información para construir el dataset, construir un modelo de análisis de regresión y luego se podrá realizar

predicciones para valores desconocidos (AWS, 2024). Este modelo estadístico usa la función logit, que es una función sigmoidea:

$$f(x) = \frac{1}{1 - e^{-x}} \tag{2}$$

Donde x es la variable independiente (conjunto de factores que inciden) y $f(x)$ la variable dependiente (probabilidad de que un hecho ocurra), la función tiene la siguiente gráfica

Figura 1.
Regresión logística



C. Machine Learning

El Machine Learning (ML) es una subcategoría de la inteligencia artificial (IA) basada en el análisis y procesamiento de datos a través de algoritmos de aprendizaje estadístico para analizar una gran cantidad de datos para identificar patrones (UC Berkeley School of Information, 2020).

Existen tipos de modelos de ML según la intervención humana o no al recolectar los datos. Así el aprendizaje será supervisado si el dataset es etiquetado y clasificado de manera previa; no supervisado cuando el reconocimiento de patrones se da sin intervención alguna del usuario e híbrido o semi-supervisado si el dataset contiene datos estructurados y no estructurados.

Los algoritmos de ML estos consideran tres componentes:

- Un proceso de decisión que recibe el resultado del procesamiento de los datos para encontrar un patrón que le permiten implementar una predicción o clasificación.
- Una función de error para validar la exactitud de la predicción, podría compararlo con ejemplos conocidos si están disponibles.
- Un proceso de optimización para el ajuste del modelo a nuevos datos de entrenamiento en una forma iterativa hasta alcanzar un nivel de precisión deseado, que permite formular una decisión final.

Dada sus características ML tiene aplicaciones en el reconocimiento de voz, chatbots, visión artificial, motores de recomendación, automatización robótica de procesos, negociación de acciones automatizada, detección de fraudes, entre otros (IBM, 2024) .

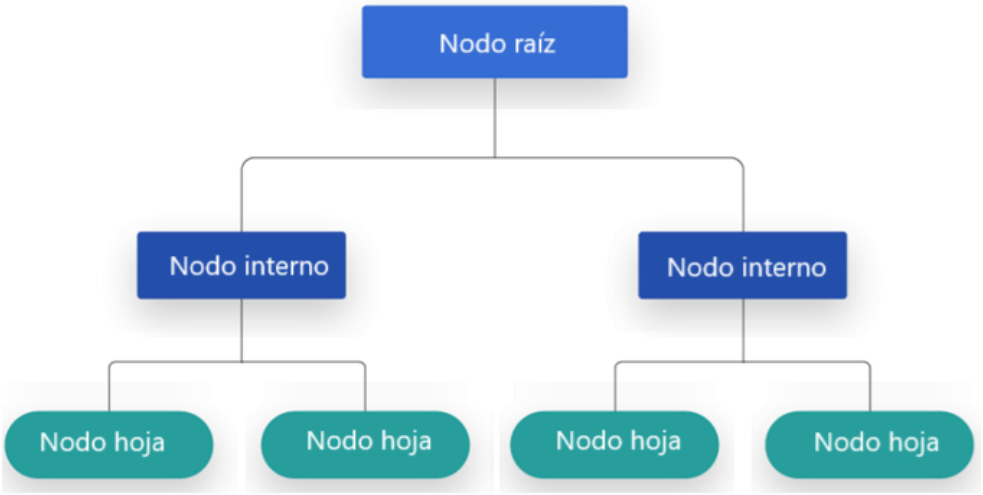
D. *Métodos de Machine Learning*

Los métodos de Machine Learning han revolucionado el análisis predictivo en finanzas, proporcionando herramientas más sofisticadas para evaluar la insolvencia. A continuación, se describen los métodos a ser utilizados en la presente investigación:

- Árboles de Decisión: este algoritmo de aprendizaje supervisado y no paramétrico se emplea en problemas de clasificación y regresión. Utiliza una estrategia de búsqueda codiciosa para determinar los puntos de división más adecuados. Su estructura jerárquica consta de un nodo raíz, nodos internos (también llamados de decisión) y nodos hoja. Los nodos internos examinan las características de los datos para dividirlos en subconjuntos más homogéneos, permitiendo una clasificación iterativa. Finalmente, los nodos hoja representan las posibles salidas del modelo para el conjunto de datos analizado.

Los árboles más pequeños tienden a generar nodos hoja más homogéneos, mientras que los árboles más grandes pueden experimentar sobreajuste y fragmentación. Para mitigar estos problemas, se aplica la poda para eliminar ramas irrelevantes, y se utiliza la validación cruzada para evaluar la eficacia del modelo (IBM, 2021).

Figura 2.
Árbol de decisión



Existen varios tipos de algoritmos de árboles de decisión, entre ellos: ID3 (Iterative Dichotomiser 3), que emplea la entropía y la ganancia de información para determinar los puntos de división; C4.5, que basa sus decisiones en la ganancia de información o la proporción de dicha ganancia; y CART (Classification and Regression Trees), que utiliza la impureza de Gini como criterio para evaluar la frecuencia de errores en la clasificación.

Estos atributos presentan las siguientes definiciones: la entropía es una medida de la impureza de los valores de la muestra (si S es el conjunto de datos, c las clases en S, p(c) proporción de datos que pertenecen a la clase c y C son todas las clases), se tiene:

$$Entropia(S) = -\sum_{c \in C} p(c) \log_2 p(c) \tag{3}$$

La ganancia de información es la diferencia en la entropía antes y después de una división para un atributo dado. El atributo con mayor ganancia de información producirá la mejor división (si a es un atributo, Sv es un conjunto de datos de S), se tiene:

$$Ganancia(a,S) = Entropia(S) - \sum_{v \in C} \frac{|Sv|}{|S|} Entropia(Sv)$$

La impureza de Gini es la probabilidad de clasificar incorrectamente un punto de datos aleatorio

$$\text{Impureza\_Gini} = 1 - \sum_{i} (p_i)^2$$

Los árboles de decisión son de fácil implementación, no requieren preparación del dataset y son flexibles; como desventaja son propensos al sobreajuste, los estimadores que genera son de alta varianza, y tienen mayor costo.

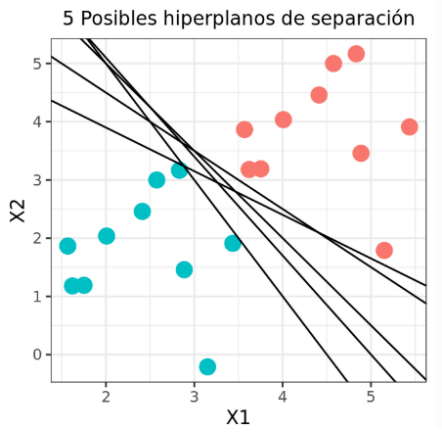
- Máquinas de Soporte Vectorial (SVM): este algoritmo se aplica tanto en clasificación como en regresión, priorizando la precisión de las predicciones y reduciendo el riesgo de sobreajuste. Su funcionamiento se basa en proyectar los datos en un espacio de características de mayor dimensión, permitiendo la separación de puntos que no son linealmente distinguibles. SVM determina un hiperplano óptimo que maximiza la distancia entre las distintas clases, conocido como margen. Gracias a esta metodología, es capaz de manejar tanto problemas lineales como no lineales mediante la transformación de los datos. Una vez entrenado el modelo, puede clasificar nuevos registros con base en los patrones aprendidos. (IBM, 2021)

  El hiperplano, en un espacio p-dimensional es un separador de dimensión p-1, con ecuación:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = 0 \qquad (4)$$

$x = (x_1, x_2, .., x_p)$ es un punto del hiperplano si cumple (4) o está a un lado o al otro si es mayor o menor a cero, por lo que el hiperplano es un clasificador. Hay conjuntos de datos que son perfectamente separables usando hiperplanos, pero se debe seleccionar el hiperplano óptimo, que será aquel que exhiba la mayor de las menores distancias de los datos con respecto a él y se le denomina el hiperplano óptimo de separación (maximal margin hyperplane), difícil de hallar cuando los datos no son perfectamente separables.

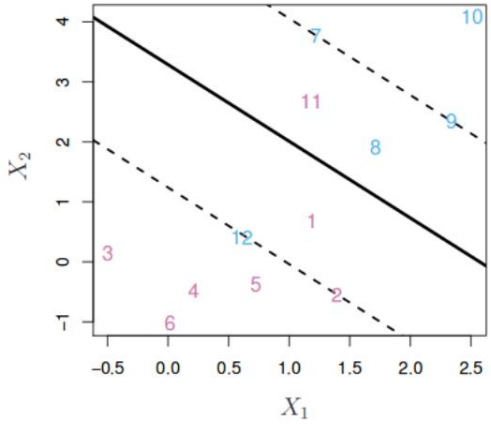
Figura 3.
Hiperplanos en un espacio bi-dimensional



La aplicación práctica son los clasificadores de vector soporte SVM, donde se permiten distancias menores a los márgenes máximos descritos, que admite la clasificación de todas las observaciones menos un número finito de ellos C, si C es infinito no se permiten violaciones, si C es cero no se sancionan las violaciones, así que el algoritmo busca minimizar C (Amat Rodrigo, 2017).

Figura 4.

Clasificador vector soporte



- Redes Neuronales Artificiales (ANN): es un modelo de aprendizaje automático inspirado en el cerebro humano, diseñado para reconocer patrones y facilitar la toma de decisiones. Su estructura está compuesta por múltiples capas de unidades interconectadas, conocidas como neuronas artificiales, que incluyen una capa de entrada, varias capas ocultas y una capa de salida. Cada una de estas neuronas posee conexiones con pesos asignados y umbrales específicos; cuando la señal procesada supera dicho umbral, la neurona se activa y transmite la información a la siguiente capa.

  A través del entrenamiento con conjuntos de datos, la red ajusta sus parámetros, optimizando su precisión conforme avanza el aprendizaje. (IBM, 2022)

  Este tipo de enfoque se enmarca dentro del aprendizaje profundo, permitiendo modelar y comprender relaciones complejas y no lineales entre los datos de entrada y salida. Su capacidad para procesar información no estructurada y extraer patrones sin un entrenamiento explícito lo hace ideal para aplicaciones en visión artificial, reconocimiento de voz y procesamiento del lenguaje natural. En términos generales, su arquitectura consta de tres niveles: una capa de entrada, múltiples capas ocultas y una capa de salida. Cada una de estas capas transforma los datos recibidos de la anterior

mediante una matriz de pesos ajustada durante el entrenamiento. Además, cada capa está formada por numerosas neuronas, que en algunos casos pueden contarse por miles.

Existen varios tipos de redes neuronales artificiales (RNA): las alimentadas hacia adelante, que son unidireccionales, donde cada neurona de una capa está conectada con todas las neuronas de la capa anterior; las de retropropagación, que ajustan su aprendizaje mediante bucles de retroalimentación correctivos, permitiendo que los datos fluyan desde el nodo de entrada hasta el de salida a través de múltiples rutas, aunque solo una es correcta, determinada por los pesos ajustados durante el entrenamiento; y las redes neuronales convolucionales, donde las capas ocultas realizan funciones de síntesis o filtrado, conocidas como convoluciones.

El aprendizaje en las RNA es supervisado, lo que significa que se ingresan conjuntos de datos etiquetados que ya contienen la respuesta correcta. La red neuronal "aprende" a partir de estas etiquetas y, una vez entrenada, es capaz de clasificar datos nuevos de manera precisa (AWS, 2024).

- Bosques Aleatorios (Random Forestes un algoritmo de aprendizaje automático supervisado desarrollado por Leo Breiman y Adele Cutler, que combina las predicciones de múltiples árboles de decisión para llegar a una conclusión general. Su funcionamiento se basa en la técnica de bagging, que emplea muestreo con reemplazo y aleatoriedad en la selección de características, con el fin de crear un conjunto de árboles de decisión que sean independientes entre sí. (IBM, 2021)

  El método de bagging consta de tres etapas: primero se divide el dataset en varios subconjuntos de forma aleatoria; luego en cada subconjunto se entrena un modelo, por lo que se generan tantos modelos como subconjuntos se hallan formado; finalmente, se combinan todos los resultados de cada uno de los modelos para obtener el resultado final, este

modelo global exhibe la robustez que no necesariamente está presente en cada modelo individual.

Random Forest se utiliza para resolver problemas de clasificación y regresión, como ventajas, funciona bien aun con datos incompletos, cada árbol funciona independientemente; aunque en contraposición exige el esfuerzo de implementar varios árboles.

2.2.3 Marco conceptual.

A. Insolvencia empresarial

Es la incapacidad de una empresa para cumplir con sus obligaciones financieras a tiempo, generalmente debido a una gestión financiera inadecuada. Existen dos tipos principales: la insolvencia de flujo de caja y la insolvencia de balance. En el caso de la insolvencia de caja, esta es una condición temporal y se presenta cuando no hay suficiente efectivo para pagar deudas inmediatas; sin embargo, ya que los activos como bienes, ahorros, inversiones u otros superan a los pasivos o deudas; es factible utilizar los activos para conseguir la liquidez requerida. En el cado de la insolvencia de balance, más grave, se da cuando las deudas o las obligaciones financieras superan el valor de los activos, haciendo imposible el pago de las mismas, en este cado de acuerdo a la legislación se podrá reestructurar las deudas, aumentar el capital u obtener financiamiento externo. Las consecuencias generadas por la insolvencia pueden ser inmediatas al afectar el flujo de caja, pero también incluyen daños al historial crediticio, dificultades futuras para acceder a créditos y, en casos extremos de insolvencia punible, sanciones legales por evadir el pago de manera ilegal. Para evitar caer en esta situación, es crucial llevar una adecuada planificación financiera y mantener un equilibrio entre ingresos y gastos (Banco Santander, 2023).

B. Variables macroeconómicas

Las Variables Macroeconómicas se definen como indicadores económicos de naturaleza cuantitativa, que proporcionan información general sobre la economía, la evaluación de las mismas permite predecir el futuro económico y la toma de decisiones en este ámbito, suelen recibir también el nombre de variables agregadas (Duana Dávila, Dueñas Soto, & Pérez Herrera, 2023). Dentro de las principales variables macroeconómicas tenemos:

- Inflación: Se trata del aumento continuo y generalizado en los precios de bienes y servicios dentro de un país a lo largo de un periodo extendido, usualmente un año. Como consecuencia, el valor de cada unidad monetaria disminuye en términos de la cantidad de productos y servicios que puede adquirir, reflejando una pérdida en el poder de compra. (Economista, 2022).
- Tasa de interés: se refiere al costo que implica obtener un préstamo o la remuneración por mantener dinero ahorrado. Se expresa como un porcentaje del capital otorgado en un crédito bancario o del monto depositado en una cuenta de ahorro. (BBVA, 2016)
- Producto bruto interno (PBI): refleja el valor total de los bienes y servicios finales producidos en un país durante un período determinado. El término "producto" hace alusión al valor añadido en la producción, mientras que "interno" señala que solo se incluye lo generado dentro del país. Además, "bruto" indica que no se consideran los cambios en los inventarios ni las variaciones en la depreciación o apreciación del capital. (MEF, 2015).
- Tasa de desempleo: evalúa la relación existente entre el nivel de desocupación en comparación con la población activa. Esta relación en su interpretación representa el porcentaje de personas en edad y condiciones adecuadas para trabajar pero que no están empleadas (Vázquez Burguillo, 2018)

C. Variables microfinancieras

Son aquellas variables que cuantifican el comportamiento de un determinado agente económico. En esta investigación consideramos:

- Solvencia: se refiere a la habilidad de una organización para satisfacer sus responsabilidades financieras a largo plazo. Se evalúa comparando el total de activos con el total de pasivos, lo que permite determinar si la entidad cuenta con el patrimonio suficiente para saldar sus deudas. (BBVA, 2016)
- Calidad de Activos: mide la proporción de activos que pueden generar ingresos y su capacidad para ser convertidos en efectivo. Un nivel elevado de activos de calidad indica que la institución posee una cartera de préstamos robusta, lo que reduce el riesgo de impagos. (BBVA, 2016)
- Eficiencia: evalúa la relación entre los costos operativos y los ingresos que genera la entidad. Una mayor eficiencia significa que la organización puede obtener más ingresos con menos gastos, resultando en una rentabilidad superior.
- Rentabilidad: Hace referencia a la capacidad de una entidad para obtener ganancias en proporción a sus ingresos, recursos o capital. Se mide a través de diversos indicadores, como el retorno sobre activos (ROA) y el retorno sobre el patrimonio. (BBVA, 2016)
- Liquidez: describe la capacidad de una organización para afrontar sus compromisos financieros a corto plazo sin incurrir en pérdidas considerables. Se evalúa mediante la proporción entre los activos líquidos, como el efectivo y sus equivalentes, y las obligaciones de corto plazo, reflejando así la disponibilidad de recursos para cubrir pagos inmediatos. (BBVA, 2016)

2.3. Definición de términos básicos

- Caja municipal de ahorro y crédito (CMAC): es una institución de microfinanciamiento (IMF) que por definición pertenece a un gobierno municipal en forma mayoritaria. En el Perú existen 12 cajas que gestionan

el 40,6% del sector de las microfinanzas. Atienden a segmentos de la población urbana que históricamente no han recibido apoyo de la banca formal. Su oferta principal es otorgar créditos destinados a microempresas y pequeñas empresas, además ofrecen cuentas y otros instrumentos de ahorro y otros servicios bancarios (Grupo de Análisis para el Desarrollo, 2021).

- Caja rural: entidades financieras que se centran en atender a segmentos de la población rural que históricamente han estado desatendidos por la banca formal. Su oferta principal incluye créditos destinados a agricultores y empresarios del sector agropecuario, así como productos de ahorro (Rebajatuscuentas, 2023).
- Machine Learning (ML): definida como una rama de la inteligencia artificial, es el conjunto de métodos y algoritmos que permiten a las computadoras analizar datos, identificar patrones y generar predicciones o decisiones sin necesidad de una programación específica para cada función, debido a que imitan la forma en la que los humanos aprenden, mejorando de forma gradual su precisión.
- Predicción: Proceso de estimar el valor futuro de una variable o resultado basado en datos históricos y patrones identificados mediante análisis estadístico o algoritmos de Machine Learning.

Capítulo III: Hipótesis y variables

3.1. Hipótesis

3.1.1. Hipótesis general.

Existe un modelo de Machine Learning que permite identificar un modelo predictivo que anticipa la insolvencia en las cajas municipales y rurales del Perú.

3.1.2. Hipótesis específicas.

- Las variables macroeconómicas y microfinancieras permiten predecir con precisión la insolvencia en las cajas municipales y rurales del Perú, utilizando técnicas de Machine Learning.
- La evaluación de los diferentes algoritmos de Machine Learning permite seleccionar cuál de ellos tiene un rendimiento predictivo superior para identificar la insolvencia en las cajas municipales y rurales del Perú
- Los métodos de Machine Learning presentan un mayor nivel de precisión y capacidad predictiva que los métodos tradicionales en la identificación de insolvencia en las cajas municipales y rurales del Perú.

3.2. Operacionalización de variables

3.2.1 Variable Dependiente.

Insolvencia en las cajas municipales y rurales del Perú.

3.2.2 Variables Independientes.

Variables macroeconómicas, que considera las dimensiones de:

- Inflación
- Tasa de interés
- Crecimiento del PBI
- Tasa de desempleo

Variables microfinancieras que considera las dimensiones de:

- Calidad de activos
- Eficiencia
- Rentabilidad
- Liquidez
- Posición de ME

3.3. Matriz de operacionalización de variables

Tabla 1.
Matriz de operacionalización de variables

VARIABLE	DEFINICIÓN CONCEPTUAL	DEFINICIÓN OPERACIONAL	DIMENSIÓN	INDICADORES	ESCALA DE VALORACIÓN	INSTRUMENTOS
Insolvencia en las cajas municipales y rurales del Perú	Capacidad de una entidad para cumplir con sus obligaciones a largo plazo.	Ratio de solvencia de las cajas municipales y rurales.	Solvencia	Ratio de Capital Global	Porcentaje (%)	Análisis financiero.
Variables macro económicas	Son un tipo de variables económicas que reflejan el comportamiento de un fenómeno macroeconómico	Variables que proporcionan información sobre el comportamiento en de la economía peruana	Inflación	Índice de Precios al Consumidor (IPC)	Porcentaje (%)	Informe de estadísticas del BCRP
			Tasa de interés	Tasa de interés promedio	Porcentaje (%)	Informe de estadísticas del BCRP
			Crecimiento del PBI	Ratio de crecimiento del PBI	Porcentaje (%)	Informe de estadísticas del BCRP
			Tasa de desempleo	Tasa de desempleo	Porcentaje (%)	Informe de estadísticas del BCRP
Variables micro financieras	Son factores directamente relacionados con el desempeño y la gestión financiera de una entidad microfinanciera,	Variables que proporcionan información clave sobre la estabilidad y sostenibilidad de las cajas	Calidad de activos	Créditos Atrasados (criterio SBS) / Créditos Directos	Porcentaje (%)	Análisis financiero.
			Eficiencia	Gastos de Administración Anualizados1 / Activo Productivo Promedio	Porcentaje (%)	Estado financiero
			Rentabilidad	Utilidad Neta Anualizada / Patrimonio Promedio (ROAE)	Porcentaje (%)	Balance general y Estado de resultados
			Liquidez	Ratio de Liquidez MN (Promedio de saldos del mes)	Porcentaje (%)	Análisis financiero
			Posición ME	Posición Global Promedio / Patrimonio Efectivo	Porcentaje (%)	Análisis financiero

Capítulo IV: Metodología del estudio

4.1. Enfoque, tipo y alcance de investigación

4.1.1 Enfoque.

El enfoque de la presente investigación es cuantitativo, debido a que su objetivo principal es analizar de manera objetiva diferentes variables macroeconómicas y microfinancieras para poder predecir la insolvencia en las cajas municipales y rurales en el Perú. Este enfoque, permite primero cuantificar las variables para posteriormente establecer relaciones estadísticas entre las mismas.

Este enfoque es el más adecuado para estudios que buscan identificar patrones y determinar relaciones entre diferentes, según (Hernandez, 2014). La elección de este enfoque, también se justifica por su capacidad de utilizar técnicas de Machine Learning, donde son necesarios un gran volumen de datos históricos para poder desarrollar modelos predictivos más precisos y confiables.

4.1.2 Tipo y alcance.

La presente investigación es de tipo explicativa, donde más allá de determinar la relación entre las variables macroeconómicas y microfinancieras, busca entender y explicar las razones detrás de estas relaciones. Usando técnicas de Machine Learning, se identifica un modelo predictivo sólido y se aborda el fenómeno en cuestión mediante un entendimiento completo y extenso. Así, será posible identificar los factores clave y desarrollar estrategias efectivas que permitan combatir o reducir los riesgos de insolvencia en las cajas municipales y rurales del Perú.

4.2. Diseño de la investigación

El diseño de esta investigación es no experimental longitudinal, ya que no se manipulan las variables, sino que se observan en su contexto natural. Se analizan

datos históricos de diez años (2013–2023) para identificar patrones y tendencias en variables macroeconómicas y microfinancieras que influyen en la insolvencia de las cajas municipales y rurales del Perú. Este enfoque permite desarrollar un modelo predictivo robusto al considerar la evolución temporal de los datos. La elección del diseño se justifica por la naturaleza observacional del estudio y el análisis dinámico de información a lo largo del tiempo.

4.3. Población y muestra

4.3.1 Población.

La población del estudio comprende todas las cajas municipales y rurales del Perú que operan bajo la supervisión de la Superintendencia de Banca, Seguros y AFP (SBS). Estas instituciones desempeñan un papel clave en la promoción de la inclusión financiera, facilitando el acceso a servicios financieros en sectores con menor desarrollo económico. Para este análisis, se consideran datos históricos de aquellas entidades que han estado en funcionamiento durante las últimas dos décadas, con un total de 2,988 registros dentro del universo de estudio.

4.3.2 Muestra.

Se empleó un muestreo no probabilístico de tipo por conveniencia. La muestra está compuesta por aquellas cajas municipales y rurales que han mantenido operaciones continuas durante el periodo 2013–2023 y que cuentan con información financiera completa y de libre acceso. Como resultado, se seleccionaron 21 cajas en total, 13 cajas municipales y 8 cajas rurales, lo que dio lugar a un total de 2,643 registros utilizados para el análisis.

4.4. Técnicas e instrumentos de recolección de datos

4.4.1 Técnicas e instrumentos.

Para esta investigación se utilizó como técnica la revisión de fuentes secundarias y como instrumento se utilizó el diccionario de variables.

Los datos secundarios relacionados con las variables macroeconómicas se obtuvieron de las estadísticas e informes proporcionados por el Banco Central de Reserva del Perú (BCRP).
Link: https://estadisticas.bcrp.gob.pe/estadisticas/series/

En cuanto a las variables microfinancieras, se utilizaron los informes financieros anuales y trimestrales de las cajas rurales y municipales, disponibles públicamente a través del portal de la Superintendencia de Banca, Seguros y AFP (SBS).
Link: https://www.sbs.gob.pe/estadisticas-y-publicaciones/estadisticas-/sistema-financiero_

4.4.2 Validez y confiabilidad.

La validez se garantizó mediante el uso exclusivo de fuentes oficiales, como la SBS y el BCRP.

Para garantizar la confiabilidad de los datos, se realizó un análisis exploratorio con el fin de detectar valores atípicos, validar rangos y verificar la consistencia general de la información.

4.4.3 Procedimiento de recolección de datos.

El proceso de recolección de datos se llevó a cabo en tres etapas:

- Identificación de instituciones: Se determinaron cuáles son las cajas rurales y municipales que han estado operativas en los últimos 10 años.
- Recolección de datos: Se realizó la extracción y organización eficiente de la información para su análisis posterior.
- Limpieza y tratamiento de los datos: Se trataron los valores faltantes y atípicos, para posteriormente normalizar y estandarizar la data para garantizar su consistencia.

Dado que los datos provienen de fuentes públicas, no fue necesario solicitar el consentimiento de las instituciones involucradas.

4.5. Técnicas de análisis de datos

El análisis de los datos se realizó utilizando técnicas de Machine Learning enfocadas en predecir la insolvencia. Se implementaron los siguientes algoritmos:

- Regresión Logística: Utilizada como modelo base, permitió estimar la probabilidad de insolvencia mediante un enfoque lineal que facilita la interpretación de los coeficientes, identificando el impacto directo de las variables macroeconómicas y microfinancieras sobre la solvencia.
- Árboles de Decisión: Este modelo generó reglas claras y jerárquicas que permiten identificar patrones clave asociados a la insolvencia. Además, su facilidad de interpretación lo hace adecuado para su aplicación práctica en el sector financiero.
- Random Forest: Modelo basado en la combinación de múltiples árboles mediante técnicas de bagging. Es capaz de manejar interacciones no lineales, reducir el sobreajuste y evaluar la importancia relativa de las variables, ofreciendo mayor precisión y robustez frente a datos desbalanceados.
- Máquinas de Vectores de Soporte (SVM): Utiliza un enfoque basado en la maximización del margen entre clases y la aplicación de kernels para capturar relaciones no lineales. SVM es especialmente útil en problemas de clasificación complejos y en la identificación de patrones sutiles en los datos.

La implementación de estos algoritmos permitió un análisis comparativo que evalúa tanto la precisión predictiva como la capacidad de cada modelo para manejar las características particulares de los datos, como el desbalance entre clases y las interacciones no lineales. Esto proporcionó un enfoque integral para seleccionar el modelo más adecuado para la predicción de insolvencia en el sector financiero peruano.

Capítulo V: Resultados

5.1. Análisis de datos

El análisis de datos se realizó utilizando el lenguaje de programación Python dentro del entorno de ejecución Google Colab, lo que permitió una ejecución eficiente del código en la nube y el acceso a bibliotecas especializadas en procesamiento de datos y aprendizaje automático.

Las principales bibliotecas utilizadas en este estudio fueron:

- Pandas y NumPy: Para la manipulación de datos y cálculos numéricos.
- Matplotlib y Seaborn: Para la visualización de resultados mediante gráficos.
- Statsmodels y SciPy: Para el análisis estadístico y pruebas de hipótesis.
- Scikit-learn: Para el entrenamiento y evaluación de modelos de regresión y clasificación.
- Imbalanced-learn (SMOTE): Para el manejo de desbalanceo en los datos.

Este enfoque permitió una ejecución robusta del análisis, asegurando precisión y eficiencia en el procesamiento de la información

5.1.1 Análisis exploratorio de los datos.

El desarrollo del modelo predictivo se inicia con un análisis exploratorio de los datos, que permite comprender las características fundamentales del conjunto de datos, identificar patrones, detectar valores atípicos y evaluar la calidad de los datos. Este paso es crucial para construir un modelo sólido y confiable.

A. Estructura del DataSet.

Como se observa en la Tabla 2. el dataset analizado contiene un total de 2655 registros distribuidos en 12 columnas. Estas variables incluyen datos categóricos, enteros y continuos. En términos de tipos de datos, una columna es de tipo *object* y almacena valores categóricos o de texto (CAJAS), dos columnas son de tipo *int64* y contienen valores enteros (AÑO y MES), mientras que las nueve columnas

restantes son de tipo *float64*, con datos numéricos continuos (PBI, INFLACION, DESEMPLEO, SOLVENCIA, CALIDAD ACTIVOS, EFICIENCIA, RENTABILIDAD, LIQUIDEZ y POSICION ME).

Tabla 2.
Estructura del DataSet

		CANTIDAD DE DATOS	TIPO DE DATOS
V. DEPENDIENTE	SOLVENCIA	2655	Float64
	CAJAS	2655	Object
	AÑO	2655	Int64
	MES	2655	Int64
	PBI	2655	Float64
	INFLACIÓN	2655	Float64
V. INDEPENDIENTE	DESEMPLEO	2655	Float64
	CALIDAD ACTIVOS	2655	Float64
	EFICIENCIA	2643	Float64
	RENTABILIDAD	2643	Float64
	LIQUIDEZ	2655	Float64
	POSICIÓN ME	2655	Float64

El análisis inicial sobre este juego de datos reveló una estructura de datos homogénea y bien definida, con una proporción limitada de valores faltantes que no comprometen la calidad general del dataset. Este diseño facilita la implementación de algoritmos de Machine Learning y asegura resultados confiables.

B. Calidad de los datos.

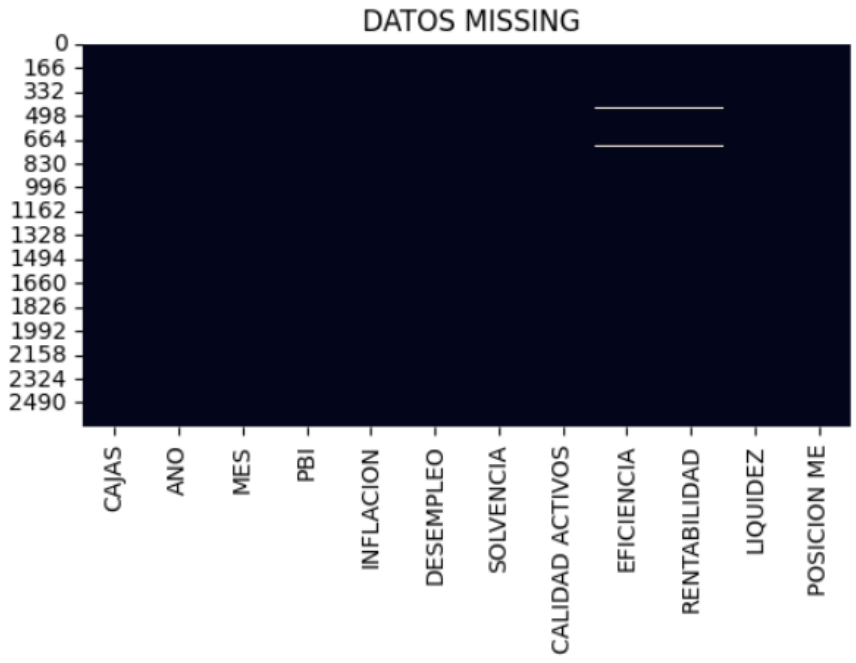
En el análisis de valores faltantes, se observó que el 99.55% de las observaciones estaban completas, mientras que dos columnas (variables EFICIENCIA y RENTABILIDAD) presentaron un total de 12 valores faltantes cada uno, equivalentes al 0.45% del total, como se muestra en la Tabla 3.

Tabla 3.
Proporción de valores faltantes

		CANTIDAD	%
V. DEPENDIENTE	SOLVENCIA	0	0
	CAJAS	0	0
	AÑO	0	0
	MES	0	0
	PBI	0	0
	INFLACIÓN	0	0
V. INDEPENDIENTE	DESEMPLEO	0	0
	CALIDAD ACTIVOS	0	0
	EFICIENCIA	12	0.45
	RENTABILIDAD	12	0.45
	LIQUIDEZ	0	0
	POSICIÓN ME	0	0

En la Figura 5 se ha resumido de manera gráfica la distribución de los valores faltantes, a partir de este resultado podemos destacar la calidad general del dataset.

Figura 5.
Gráfico de valores faltantes



La aseveración anterior proviene de observar que existe una proporción muy baja de valores faltantes, lo que permite el manejo eficiente mediante la eliminación de las filas afectadas, asegurando que la integridad y consistencia del conjunto de datos no se viera comprometida.

Adicionalmente, se realizó un proceso de limpieza y normalización, en el que se implementaron las siguientes acciones:

- Estandarización de unidades y escalas para las variables numéricas continuas.
- Revisión de valores extremos y outliers para identificar patrones atípicos y tratarlos adecuadamente.

Estas etapas son cruciales para garantizar la calidad del análisis exploratorio y el desarrollo de modelos predictivos confiables.

C. Categorización de variables.

Se realizó la categorización de las variables para mejorar la interpretabilidad, la robustez y la efectividad del análisis de datos, especialmente cuando se trabaja con grandes conjuntos de datos en los que las variables pueden tener distribuciones complejas o valores atípicos. La categorización se realizó con base en criterios utilizados por organismos internacionales como el Fondo Monetario Internacional (FMI), el Banco Mundial y la Comisión Económica para América Latina y el Caribe (CEPAL) (FMI, 2023; CEPAL, 2022; Banco Mundial, 2023).

- **SOLVENCIA**

1. Muy alto (> 15%): Indica una posición financiera sólida, con amplio margen para enfrentar pérdidas inesperadas.
2. Alto (12% - 15%): Refleja una buena gestión de riesgos y cumplimiento regulatorio.
3. Moderado (10% - 12%): Es un nivel aceptable, pero con menor capacidad para absorber grandes pérdidas.
4. Bajo (8% - 10%): Señala vulnerabilidad financiera, lo que puede preocupar a reguladores y depositantes.

5. Muy bajo (< 8%): Alto riesgo de insolvencia; puede llevar a intervención o liquidación.

- **PBI**

1. Muy alto (> 7%): Representa una expansión económica acelerada, generalmente impulsada por inversiones, exportaciones o consumo interno. Sin embargo, podría generar presiones inflacionarias o desequilibrios económicos.
2. Alto (5% - 7%): Indica un crecimiento dinámico que fomenta el empleo y mejora la calidad de vida, sin un riesgo elevado de sobrecalentamiento.
3. Moderado (2% - 5%): Es un nivel sostenible que refleja un buen equilibrio entre crecimiento e inflación. Ideal para economías maduras.
4. Bajo (0% - 2%): Señala debilidad económica, baja inversión o consumo reducido, lo que podría derivar en pérdida de empleo.
5. Negativo (< 0%): Refleja contracción económica, asociada a recesión, crisis financiera o shocks externos.

- **INFLACIÓN**

1. Muy baja (< 1%): Puede indicar deflación, una caída sostenida de precios que reduce los ingresos empresariales y ralentiza la economía.
2. Baja (1% - 3%): Representa estabilidad económica, con precios controlados que mantienen el poder adquisitivo. Ideal para atraer inversiones.
3. Moderada (3% - 6%): Suele acompañar el crecimiento económico, aunque puede afectar a consumidores con ingresos fijos.
4. Alto (6% - 10%): Genera incertidumbre, encarece bienes y servicios, y perjudica a los hogares más vulnerables.
5. Muy alta (> 10%): Señala una inflación fuera de control (hiperinflación en casos extremos), que socava la confianza en la moneda local y genera crisis económicas.

- **DESEMPLEO**

1. Muy baja (< 3%): Puede sugerir una economía muy robusta, pero con riesgo de falta de mano de obra calificada y aumento de salarios.
2. Baja (3% - 5%): Refleja alta empleabilidad y estabilidad económica, considerada saludable en la mayoría de los países.
3. Moderada (5% - 10%): Indica desafíos en la absorción de la fuerza laboral, aunque manejables.
4. Alto (10% - 15%): Señala una economía en dificultades, con altos niveles de desempleo que afectan el consumo y el crecimiento.
5. Muy alta (> 15%): Es una señal de crisis laboral severa, que requiere intervenciones gubernamentales para estimular el empleo.

- **CALIDAD DE ACTIVOS**

1. Muy bajo (< 2%): Cartera extremadamente sana, con mínimo riesgo de incumplimiento.
2. Bajo (2% - 5%): Indica una cartera saludable y bien gestionada.
3. Moderado (5% - 10%): Requiere monitoreo cercano para evitar deterioros significativos.
4. Alto (10% - 15%): Indica problemas serios de morosidad que afectan la rentabilidad.
5. Muy alto (> 15%): Alto riesgo de pérdidas que comprometen la estabilidad financiera.

- **EFICIENCIA Y GESTIÓN**

1. Muy bajo (< 30%): Altísima eficiencia operativa; gran parte del margen se convierte en beneficio neto.
2. Bajo (30% - 40%): Indica una relación favorable entre gastos y rentabilidad.
3. Moderado (40% - 50%): Gestión aceptable, aunque con margen para optimizar costos.

4. Alto (50% - 60%): Ineficiencia creciente que limita la rentabilidad.
5. Muy alto (> 60%): Signo de ineficiencia severa, con altos costos que reducen el margen de utilidad.

- **RENTABILIDAD**

1. Muy alto (> 20%): Indica una gestión eficiente y alta rentabilidad.
2. Alto (15% - 20%): Buen rendimiento financiero que atrae inversores.
3. Moderado (8% - 15%): Rentabilidad aceptable, pero con espacio para mejorar.
4. Bajo (5% - 8%): Sugiere baja eficiencia en el uso del capital.
5. Muy bajo (< 5%): Rentabilidad insuficiente; puede reflejar problemas estructurales.

- **LIQUIDEZ**

1. Muy alto (> 30%): Excelente capacidad para hacer frente a compromisos de corto plazo.
2. Alto (25% - 30%): Buena estabilidad financiera con reservas suficientes.
3. Moderado (15% - 25%): Liquidez adecuada, pero con margen para mejorar.
4. Bajo (10% - 15%): Riesgo creciente de insolvencia; se requiere mejor gestión.
5. Muy bajo (< 10%): Señal de posible crisis de liquidez, con urgencia de intervención.

- **POSICION EN MONEDA EXTRANJERA**

1. Muy bajo (< 5%): Exposición mínima; gestión prudente y conservadora.
2. Bajo (5% - 10%): Buen control de riesgos asociados al mercado cambiario.
3. Moderado (10% - 20%): Exposición controlada, aunque con riesgos latentes.
4. Alto (20% - 30%): Riesgo significativo que puede impactar negativamente la estabilidad.

5. Muy alto (> 30%): Alto vulnerabilidad a fluctuaciones de divisas, lo que podría generar pérdidas importantes.

A partir de esta categorización se implementa el análisis de estas variables mostrado en los apartados d) y e)

D. Análisis de la variable objetivo SOLVENCIA.

Como se puede observar en la Tabla 4, se analizaron 2,643 observaciones del indicador de solvencia correspondiente a cajas municipales y rurales del Perú. La media se ubicó en 14.62%, y la mediana en 14.51%, lo que indica que, en general, las entidades presentan una posición financiera saludable, dentro del rango definido como "Alto" (12% - 15%), cercano incluso al nivel "Muy Alto" (> 15%).

El percentil 25 (13.23%) y el percentil 75 (15.83%) muestran que el 50% central de las entidades se concentra entre los niveles "Alto" y "Muy Alto", lo cual sugiere una adecuada gestión de riesgos y cumplimiento regulatorio por parte de la mayoría de cajas.

Sin embargo, se observan valores extremos. El valor mínimo registrado fue de 5.13%, clasificado como "Muy Bajo", lo que implica un alto riesgo de insolvencia en al menos una entidad, representando una alerta importante para los reguladores. Asimismo, el valor máximo alcanzó 65.55%, muy por encima de los rangos definidos, lo que indica la presencia de valores atípicos positivos que deben analizarse más a fondo.

La desviación estándar de 2.89 sugiere una variabilidad moderada en los niveles de solvencia, aunque algunos valores extremos podrían estar influyendo en este resultado.

En conjunto, estos resultados muestran que, si bien la mayoría de las cajas presentan niveles sólidos de solvencia, existe un grupo reducido con posibles problemas financieros severos, lo cual justifica la necesidad de un sistema de monitoreo y predicción continuo como el propuesto en esta investigación.

Tabla 4.
Resumen descriptivo de la variable SOLVENCIA

	VALOR
CONTEO	2643.000000
MEDIA	14.617522
DESVIACIÓN	2.889687
MÍNIMO	5.130000
25%	13.225000
50%	14.510000
75%	15.835000
MÁXIMO	65.550000

Para la implementación del análisis se utilizó Python (el código está en los anexos) el resultado se muestra en tres herramientas, una tabla de distribución de frecuencias de cada categoría, una gráfica de la misma distribución en su forma de barras y el boxplot de la misma variable.

Como muestra la Tabla 5. las categorías superiores, Muy Alto (1) y Alto (2), dominan, representando en conjunto el 85.24% de las observaciones. En contraste, las categorías inferiores, Bajo (4) y Muy Bajo (5), están subrepresentadas, con menos del 3% de las observaciones combinadas.

Tabla 5.
Distribución de la variable SOLVENCIA

	CANTIDAD	%
2. ALTO (12% - 15%)	1261	47.71
1. MUY ALTO (> 15%)	992	37.53
3. MODERADO (10% - 12%)	310	11.73
4. BAJO (8% - 10%)	72	2.72
5. MUY BAJO (< 8%)	8	0.30

La prevalencia de las categorías Muy Alto y Alto refleja que la mayoría de las cajas tienen buenos indicadores de solvencia. Sin embargo, las pocas observaciones en las categorías Baja y Muy Baja generan un desbalance en la distribución de las

categorías lo cual puede afectar negativamente el desempeño de los modelos predictivos, ya que tienden a estar sesgados hacia las clases mayoritarias. Esto es especialmente problemático para modelos como la regresión logística y SVM, que asumen distribuciones más balanceadas.

Figura 6.
Gráfica de frecuencia de la variable SOLVENCIA

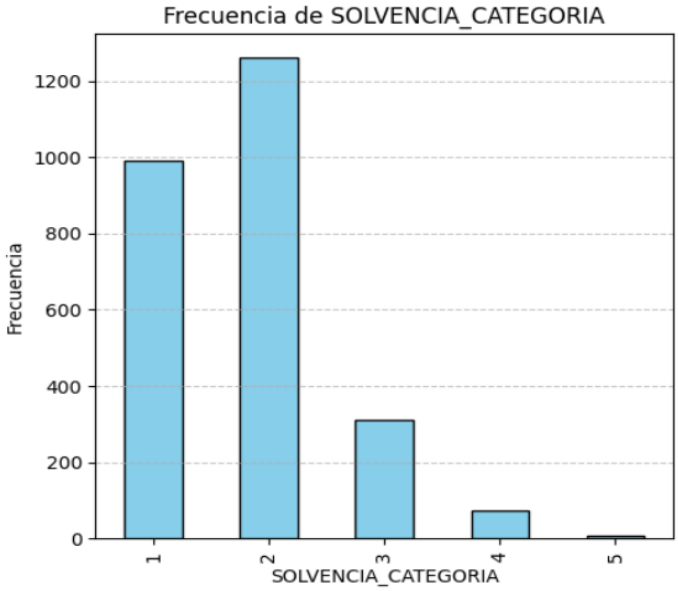
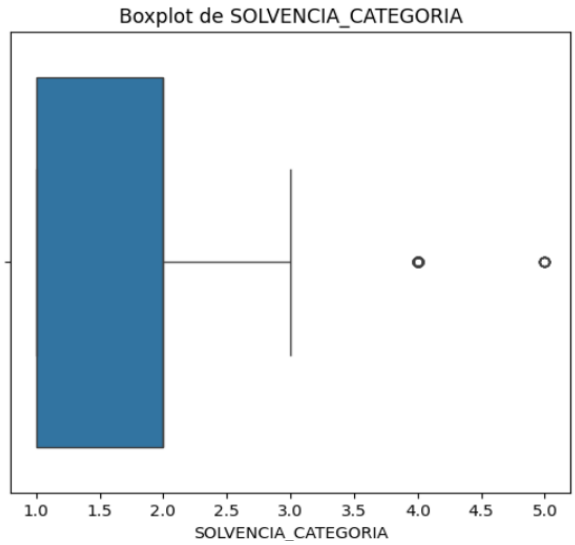


Figura 7.

Gráfico de boxplot de la variable SOLVENCIA



E. Análisis de las variables independientes.

Aquí se analizan las variables financieras restantes consignadas en el DataSet.

Para la implementación del análisis de cada variable se utilizó Python (el código está en los anexos) el resultado se muestra en tres herramientas, una tabla de distribución de frecuencias de cada categoría, una gráfica de la misma distribución en su forma de barras y el boxplot de la misma variable.

- **PBI**

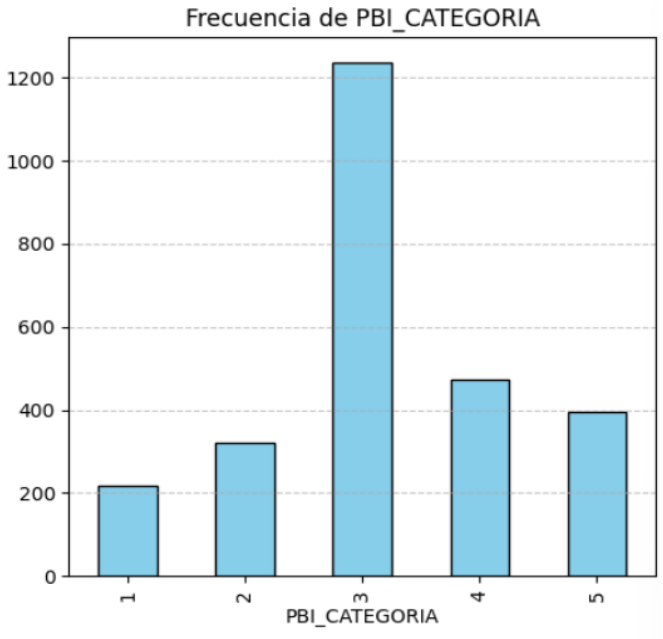
Para la variable PBI la mayoría de los registros (46.73%) se encuentran en la categoría 3 (Moderado: 2% - 5%), lo que indica que un gran porcentaje de las cajas rurales y municipales tienen un PBI moderado. Las categorías extremas (1: Muy alto y 5: Negativo) tienen baja representación, con solo el 8.21% y el 15.02%, respectivamente, lo que sugiere que las fluctuaciones extremas en el PBI son menos frecuentes.

Tabla 6.

Distribución de la variable PBI

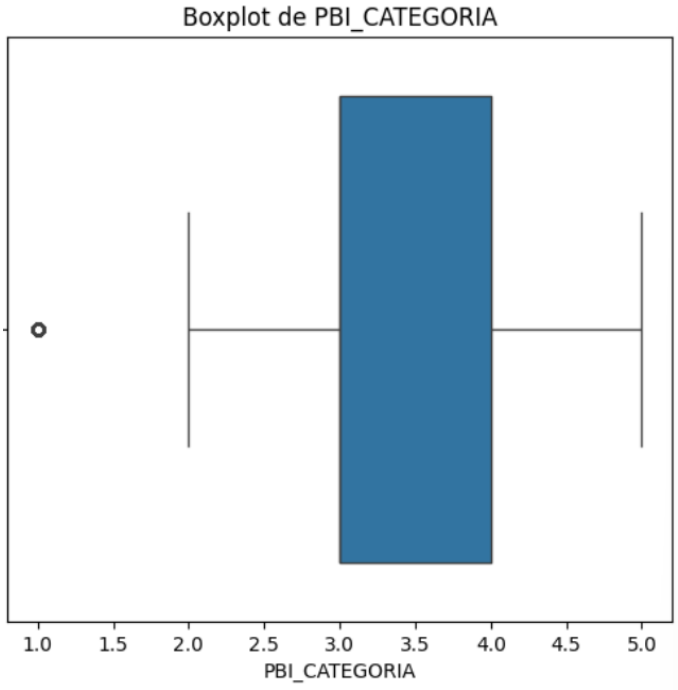
	CANTIDAD	%
3. MODERADO (2% - 5%)	1235	46.73
4. BAJO (0% - 2%)	472	17.86
5. MUY BAJO (< 0%)	397	15.02
2. ALTO (5% - 7%)	322	12.18
1. MUY ALTO (> 7%)	217	8.21

Figura 8.
Gráfico de frecuencia de la variable PBI



La variable PBI muestra rangos homogéneos, aunque existen algunos valores extremos en las categorías más bajas, lo que sugiere que las fluctuaciones en el PBI no son frecuentes.

Figura 9.
Gráfica boxplot de la variable PBI



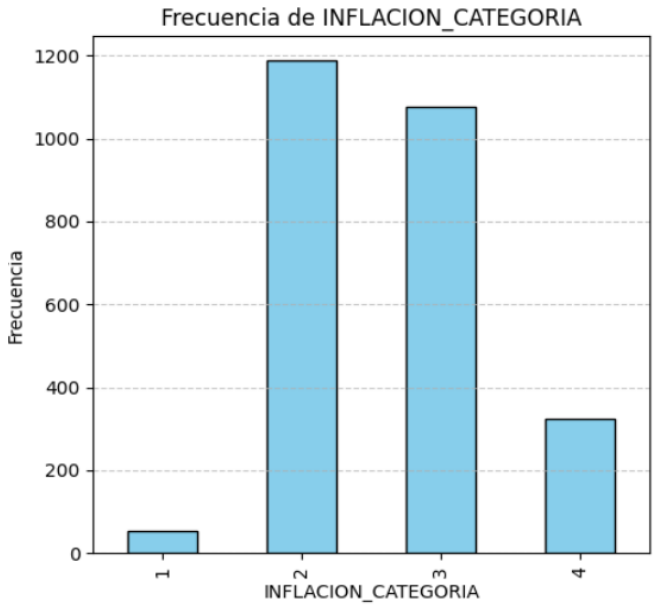
- **INFLACION**

En cuanto a la inflación, las categorías 2 (Baja: 1% - 3%) y 3 (Moderada: 3% - 6%) predominan, representando más del 85% de los registros. Esto indica que la mayoría de las cajas rurales operan en economías con niveles de inflación bajos a moderados. Las inflaciones muy altas (5) y muy bajas (1) son poco frecuentes.

Tabla 7.
Distribución de la variable INFLACION

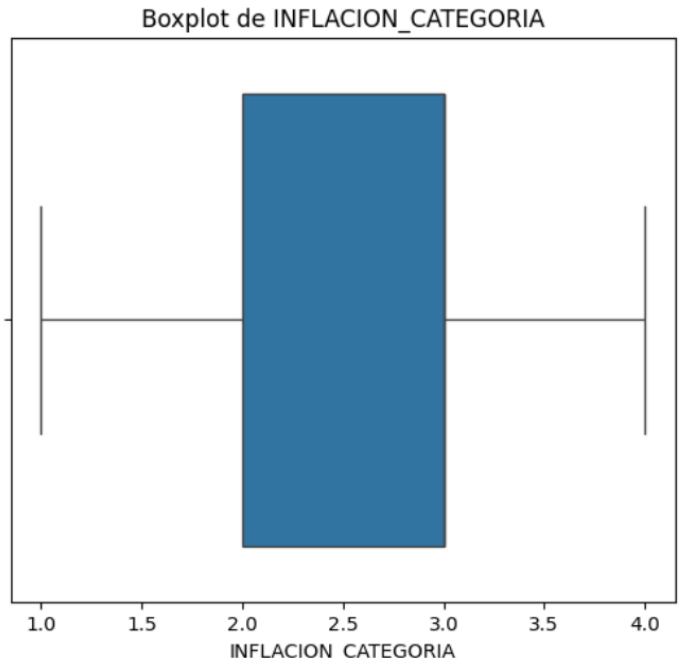
	CANTIDAD	%
2. BAJA (1% - 3%)	1188	44.95
3. MODERADA (3% - 6%)	1077	40.75
4. ALTA (6% - 10%)	324	12.26
1. MUY BAJA (< 1%)	54	2.04

Figura *10*.
Gráfica de frecuencia de la variable INFLACION



Podemos observar que la variable INFLACIÓN muestra rangos homogéneos, lo que sugiere que las fluctuaciones en la inflación no son frecuentes.

Figura 11.
Gráfica boxplot de la variable INFLACION



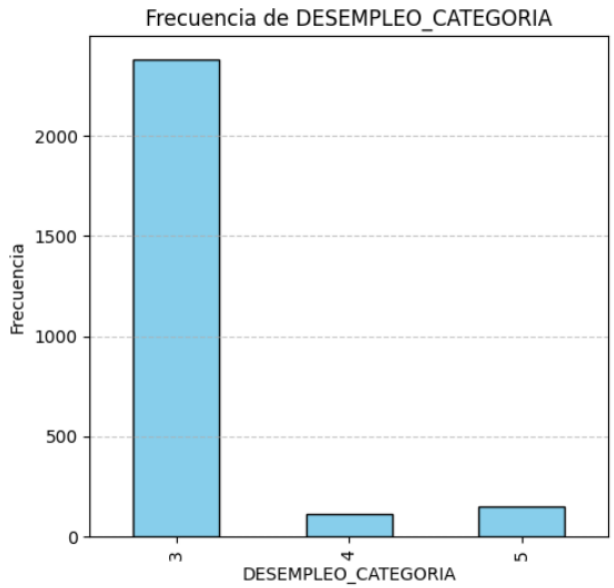
- **DESEMPLEO**

Sobre el desempleo, la categoría 3 (Moderado: 5% - 10%) es la más común, con el 90% de los registros en esta categoría. Esto refleja que la mayor parte de las cajas rurales se encuentran en regiones con tasas de desempleo moderadas. Las categorías más altas (4 y 5) tienen una representación menor, lo que sugiere que los altos niveles de desempleo son menos frecuentes.

Tabla 8.
Distribución de la variable DESEMPLEO

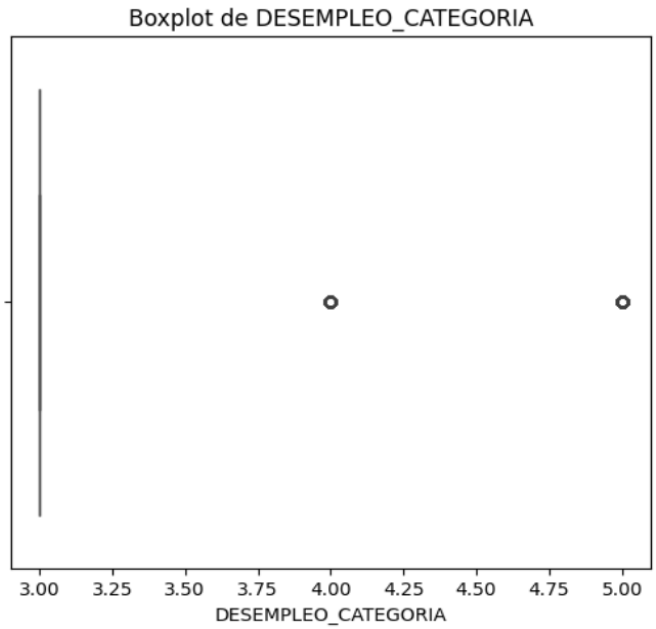
	CANTIDAD	%
3. MODERADA (5% - 10%)	2379	90.01
5. MUY ALTA (> 15%)	151	5.71
4. ALTA (10% - 15%)	113	4.28

Figura 12.
Gráfica de frecuencia de la variable DESEMPLO



Para la variable desempleo, La mayor parte de los registros se concentran en niveles moderados de desempleo (5% - 10%). Las tasas de desempleo muy altas (mayores al 15%) son menos frecuentes, lo que refleja una estabilidad en los niveles de desempleo en las regiones donde operan las cajas rurales.

Figura 13.
Gráfica boxplot de la variable DESEMPLEO



- **CALIDAD DE ACTIVOS**

Referido a la calidad de activos, la categoría 3 (Moderada: 5% - 10%) predomina, con más del 50% de los registros en esta categoría. Esto indica que la mayoría de las cajas rurales tienen una calidad de activos moderada. Las categorías extremas (1: Muy bajo y 5: Muy alto) tienen una representación combinada inferior al 15%, lo que sugiere que los extremos en la calidad de los activos son poco comunes.

Tabla 9.
Gráfica de frecuencia de la variable CALIDAD DE ACTIVOS

	CANTIDAD	%
3. MODERADO (5% - 10%)	1377	52.10
2. BAJO (2% - 5%)	587	22.21
5. MUY ALTO (> 15%)	374	14.15
4. ALTO (10% - 15%)	304	11.50
1. MUY BAJO (< 2%)	1	0.04

Figura 14.
Gráfica de frecuencia de la variable CALIDAD DE ACTIVOS

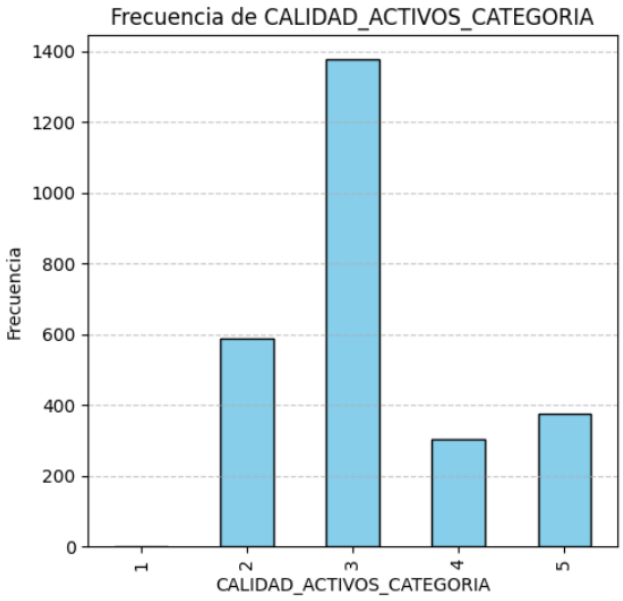
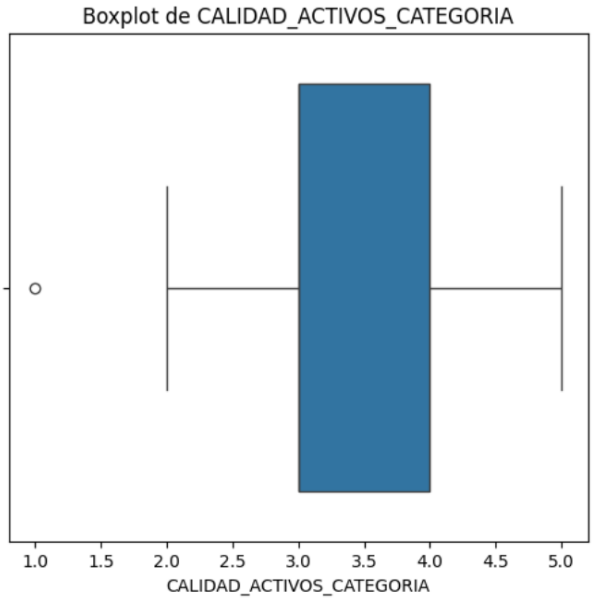


Figura 15.
Gráfica boxplot de la variable CALIDAD DE ACTIVOS



- **EFICIENCIA**

Los datos de la variable eficiencia están concentrados en los niveles más altos de eficiencia, con algunos valores atípicos en la categoría más baja. Esto refleja que la mayoría de las cajas rurales tienen una eficiencia alta, aunque hay algunos casos que se alejan significativamente del promedio.

Tabla 10.
Distribución de la variable EFICIENCIA

	CANTIDAD	%
5. MUY ALTO (> 60%)	1570	59.40
1. MUY BAJO (< 30%)	599	22.66
4. ALTO (50% - 60%)	426	16.12
2. BAJO (30% - 40%)	46	1.74
3. MODERADO (40% - 50%)	2	0.08

Figura *16*.

Gráfica de frecuencia de la variable EFICIENCIA

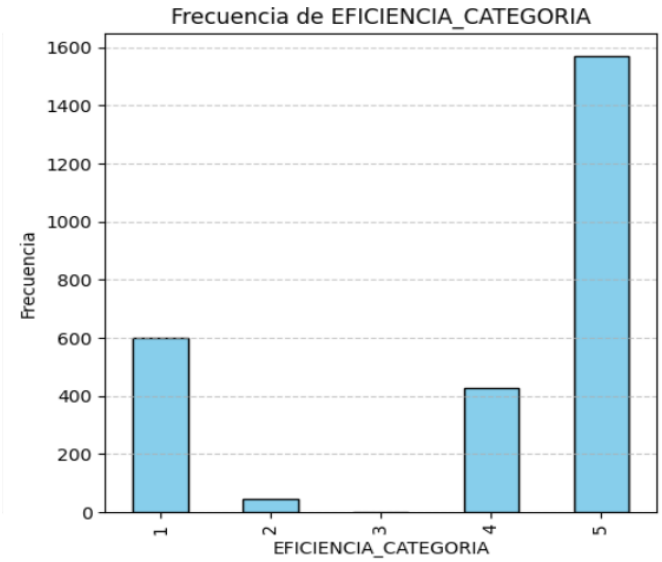
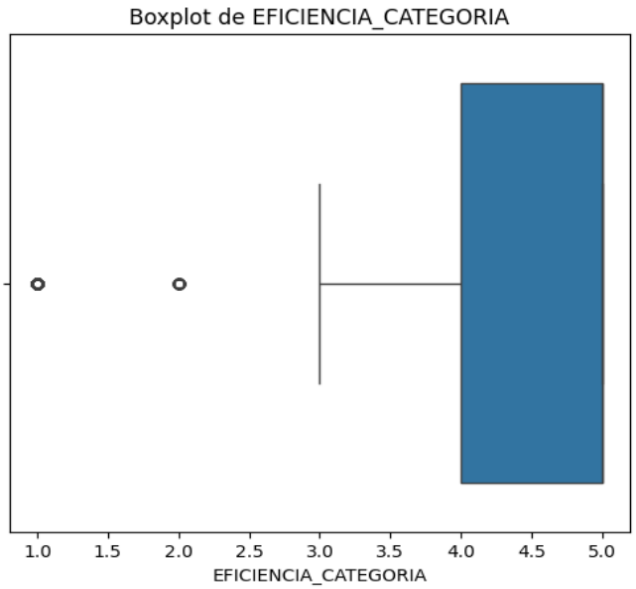


Figura *17*.

Gráfica boxplot de la variable EFICIENCIA



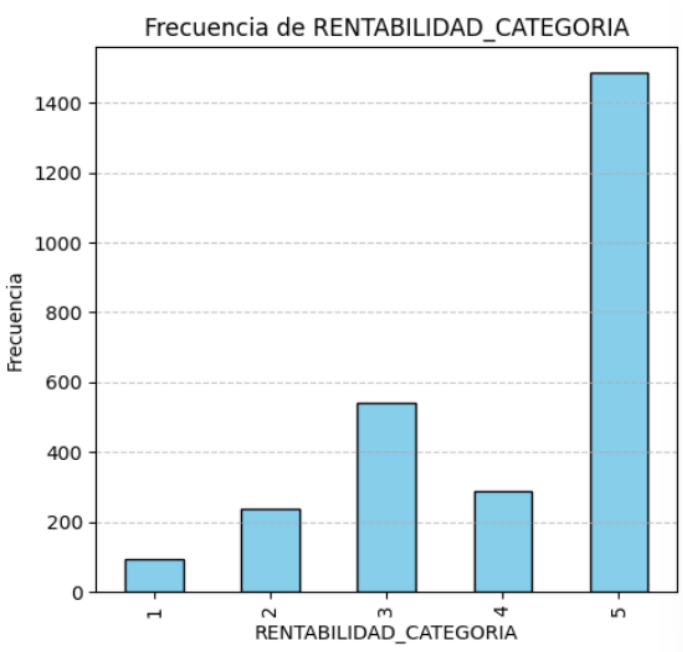
- **RENTABILIDAD**

La rentabilidad de la mayoría de las cajas rurales (56.22%) se encuentran en la categoría 5 (Muy baja: < 5%), lo que indica que una gran parte de las entidades tiene rentabilidad baja. Las categorías más altas (1 y 2) representan menos del 12%, lo que sugiere que las rentabilidades elevadas son poco frecuentes en este contexto.

Tabla 11.
Distribución de la variable RENTABILIDAD

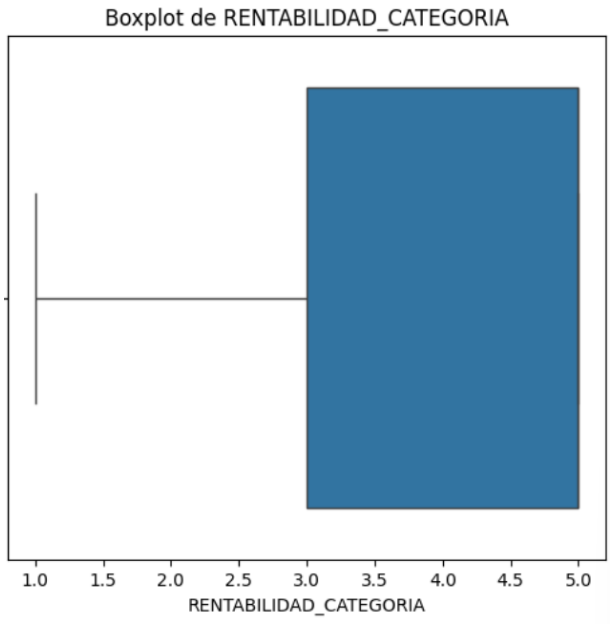
	CANTIDAD	%
5. MUY BAJO (< 5%)	1486	56.22
3. MODERADO (8% - 15%)	540	20.43
4. BAJO (5% - 8%)	287	10.86
2. ALTO (15% - 20%)	236	8.93
1. MUY ALTO (> 20%)	94	3.56

Figura *18*.
Gráfica de frecuencia de la variable RENTABILIDAD



La rentabilidad muestra una alta dispersión en las categorías superiores, lo que sugiere que hay una gran variabilidad en los niveles de rentabilidad. Esto podría indicar que algunas cajas tienen una rentabilidad significativamente diferente al resto.

Figura *19*
Gráfica boxplot de la variable RENTABILIDAD



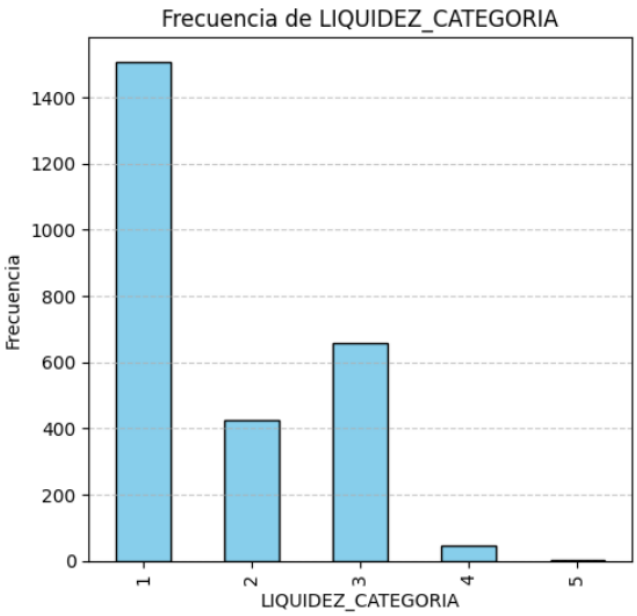
- **LIQUIDEZ**

En la variable LIQUIDEZ, la categoría 1 (Muy alta: > 30%) es la más representada, con el 57% de los registros. Esto indica que muchas cajas rurales tienen un nivel de liquidez relativamente alto. Las categorías más bajas (4 y 5) representan solo el 2% de los registros, lo que muestra que pocos casos presentan una liquidez muy baja.

Tabla 12.
Distribución de la variable LIQUIDEZ

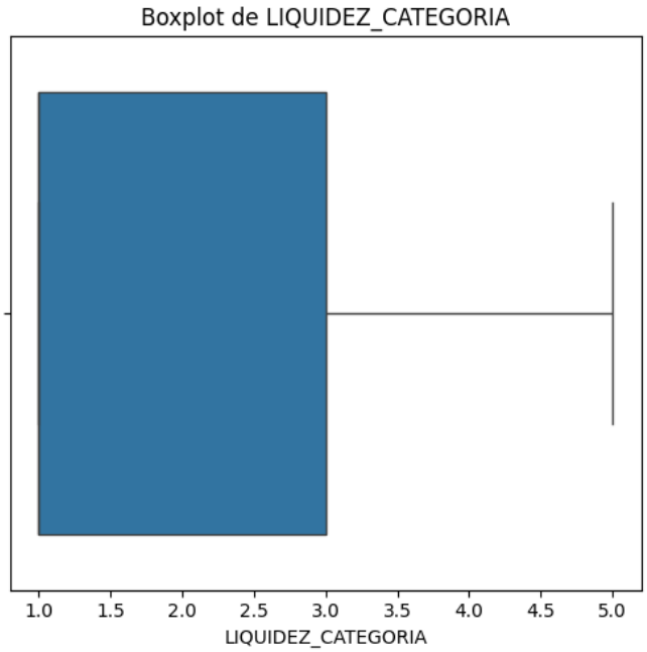
	CANTIDAD	%
1. MUY ALTO (> 30%)	1507	57.02
3. MODERADO (15% - 25%)	660	24.97
2. ALTO (25% - 30%)	427	16.16
4. BAJO (10% - 15%)	45	1,70
5. MUY BAJO (< 10%)	4	0.15

Figura 20.
Gráfica de frecuencia de la variable LIQUIDEZ



La liquidez muestra una alta dispersión en las categorías superiores, lo que sugiere que hay una gran variabilidad en los niveles de liquidez. Esto podría indicar que algunas cajas tienen una liquidez significativamente diferente al resto.

Figura 21.
Gráfica boxplot de la variable LIQUIDEZ



- **POSICIÓN DE MONEDA EXTRANJERA**

Tabla 13.
Distribución de la variable POSICIÓN DE MONEDA EXTRANJERA

	CANTIDAD	%
1. MUY BAJO (< 5%)	2490	94.21
2. BAJO (5% - 10%)	91	3.44
3. MODERADO (10% - 20%)	29	1.10
5. MUY ALTO (> 30%)	26	0.98
4. ALTO (20% - 30%)	7	0.26

Figura 22.

Gráfica de frecuencia de la variable POSICIÓN DE MONEDA EXTRANJERA

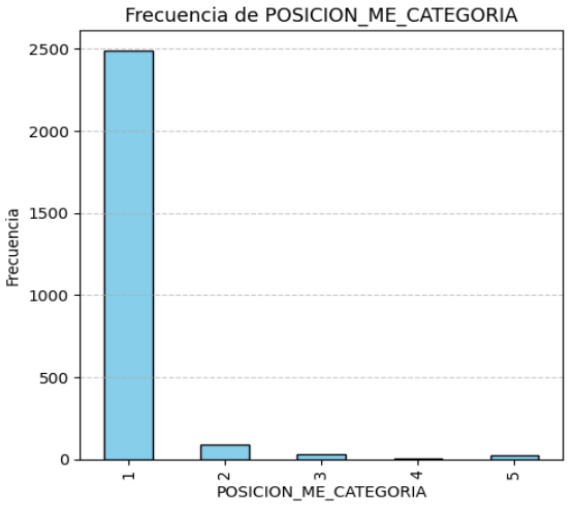
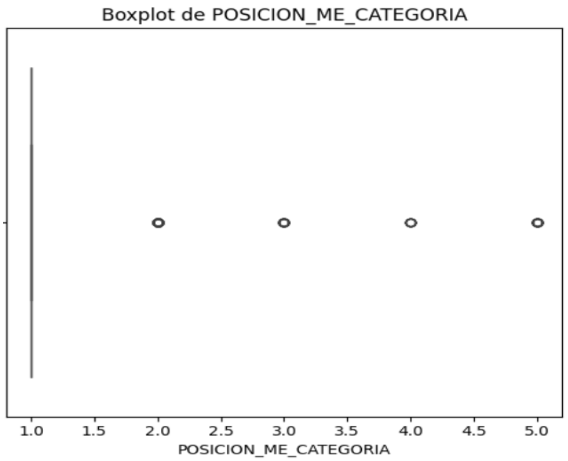


Figura 23.
Gráfica boxplot de la variable POSICIÓN DE MONEDA EXTRANJERA



5.1.2 Selección de variables.

La selección de variables relevantes es un paso crucial en la construcción del modelo predictivo. Identificar los factores que más influyen en la variable objetivo (SOLVENCIA) no solo mejora la precisión del modelo, sino que también facilita la interpretación de los resultados y su aplicación práctica. Para seleccionar las variables más relevantes, se utilizó la matriz de correlación que se muestra en la Figura 20, la que ha sido construida en Python, el código se ha consignado en los anexos de la presente tesis. En esta matriz se muestran las relaciones entre la variable objetivo y las variables predictoras seleccionadas.

Figura 24.
Matriz de correlación de todas las variables

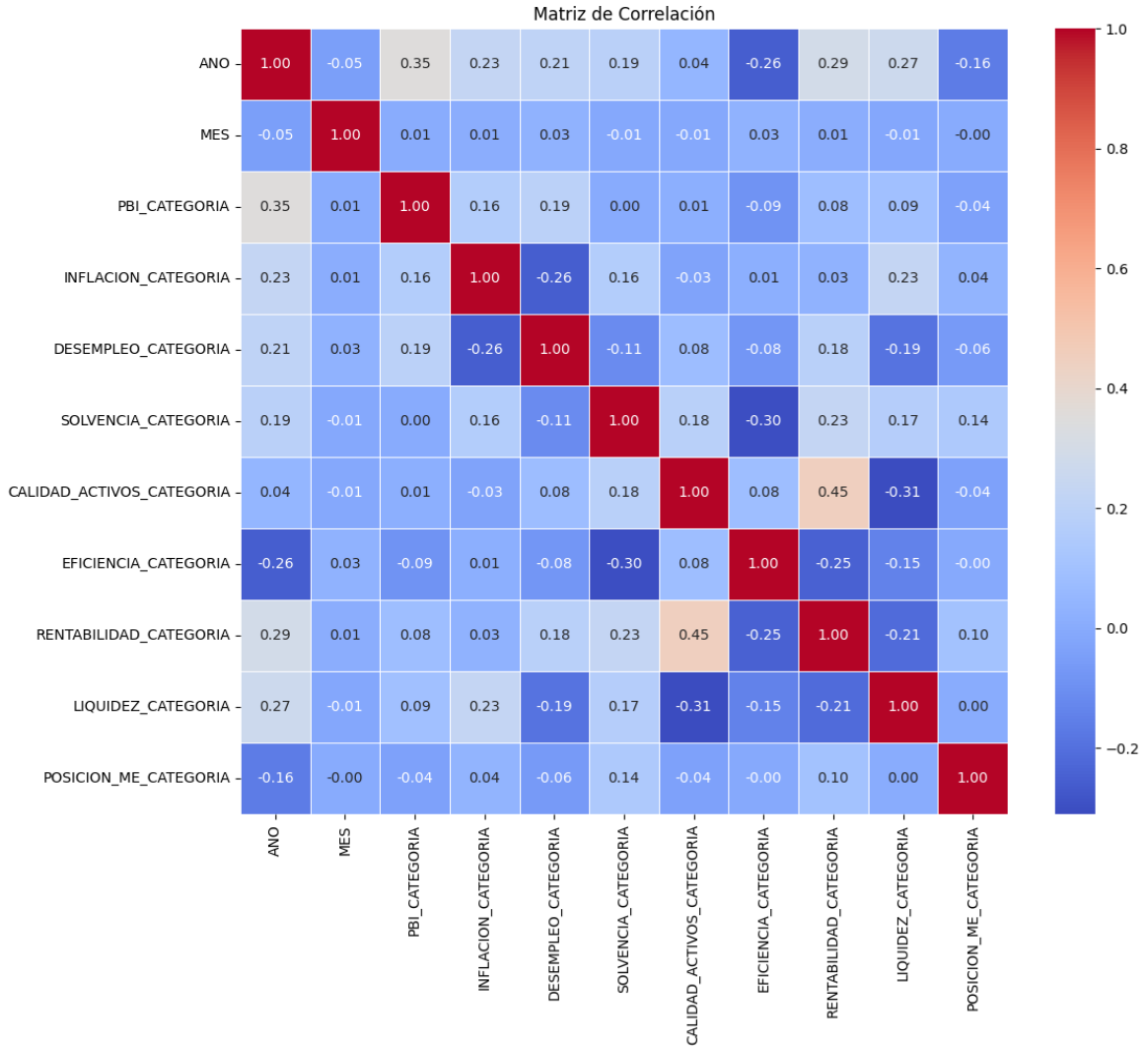


Tabla 14.

Correlación de todas las variables

	SOLVENCIA
SOLVENCIA	1.000000
RENTABILIDAD	0.227468
AÑO	0.190614
CALIDAD ACTIVOS	0.181122
LIQUIDEZ	0.166696
INFLACIÓN	0.164008
POSICIÓN ME	0.140391
PBI	0.000984
MES	-0.011486
DESEMPLEO	-0.111285
EFICIENCIA	-0.300838

A. Variables con correlación positiva.

Estas variables presentan una relación directa con la SOLVENCIA. A mayor valor de la variable, mayor es la probabilidad de tener una categoría de SOLVENCIA más alta.

- RENTABILIDAD (0.227): Esta es la variable con mayor correlación positiva con SOLVENCIA. Indica que una mayor rentabilidad está asociada con niveles más altos de solvencia, lo cual es consistente con la teoría financiera, ya que un desempeño financiero sólido suele traducirse en mayor estabilidad.
- ANO (0.191): Sugiere que la solvencia mejora ligeramente con los años. Esto podría estar relacionado con mejoras en la gestión financiera, cambios regulatorios o condiciones macroeconómicas más favorables en los años recientes.
- CALIDAD_ACTIVOS (0.181): Indica que una mejor calidad de los activos (menores niveles de cartera problemática) está asociada con una mayor solvencia.

- LIQUIDEZ (0.167): La liquidez también muestra una correlación positiva, aunque moderada, lo que implica que las cajas con más recursos líquidos tienden a estar en categorías de solvencia más altas.
- INFLACION (0.164): La relación positiva moderada podría reflejar que en períodos de inflación moderada, las cajas tienen mayor capacidad de ajuste en sus operaciones financieras, favoreciendo su solvencia.
- POSICION_ME (0.140): La posición en el mercado tiene una influencia positiva menor, indicando que las cajas con un mejor posicionamiento tienden a ser más solventes.

B. Variables con correlación negativa.

Estas variables tienen una relación inversa con la SOLVENCIA. A mayor valor de la variable, menor es la probabilidad de estar en una categoría de SOLVENCIA más alta.

- DESEMPLEO (-0.111): Indica que mayores tasas de desempleo tienden a estar asociadas con una menor solvencia. Esto es consistente con la lógica económica, ya que mayores niveles de desempleo podrían generar dificultades para los prestatarios, aumentando el riesgo crediticio.
- EFICIENCIA (-0.301): Es la variable con mayor correlación negativa. Esto sugiere que mayores niveles de gastos operativos (baja eficiencia) están fuertemente asociados con menores niveles de solvencia. Esta relación destaca la importancia de mantener una gestión operativa eficiente para preservar la estabilidad financiera.

C. Variables con correlación muy baja o nula.

Estas variables tienen un impacto casi nulo en la SOLVENCIA, según la matriz de correlación.

- PBI (0.001): La relación con el PBI es prácticamente inexistente, lo que sugiere que el desempeño económico del país no afecta directamente la solvencia de las cajas. Esto podría deberse a la naturaleza local y específica de sus operaciones.

- MES (-0.011): El mes del año no tiene un efecto relevante en la solvencia, lo que podría indicar que la estacionalidad no influye de manera significativa en las operaciones financieras de las cajas.

Con el objetivo de evaluar la posible presencia de multicolinealidad entre las variables independientes incluidas en el modelo, se procedió al cálculo del Factor de Inflación de la Varianza (VIF, por sus siglas en inglés) para cada una de ellas. Este análisis permite cuantificar el grado en que una variable está correlacionada linealmente con las demás.

Tabla 15.
Multicolinealidad de todas las variables independientes

	VIF
AÑO	1.613846
MES	1.007283
PBI	1.188495
INFLACIÓN	1.229290
DESEMPLEO	1.270074
CALIDAD ACTIVOS	1.398825
EFICIENCIA	1.185991
RENTABILIDAD	1.642526
LIQUIDEZ	1.378247
POSICIÓN ME	1.082378

Como se muestra en la Tabla 15 todos los valores de VIF se encuentran muy por debajo del umbral crítico de 5, con rangos que van desde 1.00 hasta aproximadamente 1.65. Esto indica que no existe una colinealidad significativa entre las variables numéricas consideradas. Por tanto, todas las variables independientes pueden mantenerse en el modelo sin riesgo de distorsión en la estimación de los coeficientes.

A partir del análisis correlacional y de multicolinealidad, se seleccionaron las siguientes cinco variables como las más relevantes para la construcción de los modelos predictivos, siguiente paso de este trabajo:

- RENTABILIDAD
- CALIDAD ACTIVOS
- LIQUIDEZ
- EFICIENCIA
- INFLACIÓN

Estas variables no solo muestran una relación estadística significativa con la solvencia, sino que también tienen interpretaciones prácticas claras para la gestión y evaluación de riesgos financieros. Su inclusión en los modelos permite capturar de manera integral tanto factores internos como externos que afectan la estabilidad financiera de las cajas rurales y municipales del Perú.

5.1.3 Construcción de los modelos de predicción.

Una vez seleccionadas las variables más relevantes se realiza la construcción de los modelos predictivos, cuyo objetivo es anticipar la insolvencia de las cajas rurales y municipales utilizando métodos de Machine Learning.

Para la construcción de los modelos se implementaron y compararon cuatro algoritmos:

- Regresión Logística
- Árboles de Decisión
- Random Forest
- Máquinas de Soporte Vectorial (SVM)

El desarrollo de cada modelo se hizo utilizando Python (el código se muestra en los anexos) y se basó principalmente en tres pasos:

- División de datos: El conjunto de datos fue dividido en un 80% para entrenamiento y un 20% para prueba, utilizando una separación estratificada para preservar la distribución de la variable objetivo (SOLVENCIA).

- Optimización de hiperparámetros: Se aplicaron técnicas como GridSearchCV para encontrar los parámetros óptimos que maximizan la precisión de cada modelo.
- Evaluación de desempeño: Se utilizaron las métricas precisión, recall, F1-score, y el área bajo la curva ROC (AUC) para medir la efectividad de cada modelo en clasificar correctamente las diferentes categorías de solvencia.

a) Regresión Logística

El modelo de regresión logística fue utilizado como modelo base para predecir la solvencia de las cajas rurales y municipales. A través de un ajuste de hiperparámetros utilizando GridSearchCV, se lograron identificar los mejores parámetros, lo que permitió mejorar la precisión del modelo. Los mejores hiperparámetros encontrados fueron un valor de C igual a 100, con penalty L1 (Lasso) y el solver 'saga'. Esto sugiere que la regularización L1, junto con un valor alto para C, proporcionó un ajuste adecuado al modelo.

Los datos cumplen con los supuestos fundamentales para la aplicación de la regresión logística, lo que garantiza la validez de los resultados obtenidos. La relación entre las variables predictoras y el logit de la variable dependiente es aproximadamente lineal, asegurando que el modelo pueda captar correctamente las tendencias subyacentes. Además, se verificó la independencia de las observaciones, evitando sesgos en la estimación de coeficientes. La ausencia de multicolinealidad fue confirmada mediante el análisis del Factor de Inflación de Varianza (VIF), garantizando la estabilidad de los predictores en el modelo. Asimismo, la variable dependiente está estructurada en categorías mutuamente excluyentes y exhaustivas, permitiendo una correcta clasificación. Finalmente, el tamaño muestral es adecuado para la estimación de parámetros, y se aplicó SMOTE para mitigar el impacto del desbalance de clases, mejorando la representatividad de los datos minoritarios.

El desempeño global del modelo fue moderado, con una exactitud del 61%. Esto significa que el modelo logró acertar en el 61% de las predicciones realizadas, lo que, aunque es aceptable como punto de partida, refleja un rendimiento limitado,

especialmente teniendo en cuenta que el problema involucra múltiples clases desbalanceadas. Este nivel de exactitud indica que el modelo tiene dificultades para clasificar correctamente las clases minoritarias.

Al observar la matriz de confusión en la Tabla 16, se puede ver que el modelo tuvo un buen rendimiento en las clases mayoritarias, como la Clase 5 (Solvencia Muy Alto), en la que el modelo logró predecir correctamente 249 de las 250 observaciones, con un recall de 1.00. Sin embargo, el modelo presentó grandes dificultades para predecir las clases minoritarias, especialmente las Clases 3, 4 y 5, con una predicción muy limitada. Las Clases 3 y 4, que representan categorías intermedias de solvencia, mostraron bajas tasas de predicción correcta, evidenciando que el modelo tiene problemas para diferenciar entre categorías que están cercanas en términos de características financieras.

Tabla 16.
Matriz de confusión Regresión Logística

	1	**2**	**3**	**4**	**5**
1	172	51	28	18	0
2	90	106	20	24	1
3	44	32	65	104	4
4	22	0	48	177	5
5	0	0	0	1	249

El reporte de clasificación mostrado en la Tabla 17, muestra que la Clase 5 (Solvencia Muy Alto) tuvo un excelente desempeño, con un F1-Score de 0.98 y un recall de 1.00, lo que refleja que el modelo fue excepcionalmente preciso al predecir los casos más solventes. En contraste, la Clase 3 (Solvencia Moderado) tuvo un F1-Score de 0.32 y un recall de 0.26, lo que indica que el modelo falló en identificar correctamente esta clase. Las Clases 1 (Solvencia Muy Bajo) y 2 (Solvencia Bajo) tuvieron un desempeño intermedio, con F1-Scores de 0.58 y 0.49, respectivamente, lo que refleja que, aunque el modelo fue algo efectivo en la predicción de estas clases, todavía necesita mejoras.

El macro promedio de F1-Score fue de 0.60, lo que refleja un rendimiento bajo en general, especialmente cuando se evalúan todas las clases de manera equitativa. El weighted average F1-Score, por otro lado, fue de 0.60, lo que indica que el modelo estaba sesgado hacia las clases más representadas, favoreciendo las predicciones correctas en esas categorías a expensas de las clases menos frecuentes.

Tabla 17.
Reporte de clasificación Regresión Logística

	PRECISIÓN	RECALL	F1-SCORE	SUPPORT
1	0.52	0.64	0.58	269
2	0.56	0.44	0.49	241
3	0.40	0.26	0.32	249
4	0.55	0.70	0.61	252
5	0.96	1.00	0.98	250
ACURRACY			0.61	1261
MACRO AVG	0.60	0.61	0.60	1261
WEIGHTED AVG	0.60	0.61	0.60	1261

El modelo de regresión logística fue utilizado como modelo inicial para predecir la solvencia de las cajas rurales y municipales. A pesar de su simplicidad y su valor como referencia, el modelo presentó limitaciones importantes. Estas incluyen el desbalance de clases, que sesgó las predicciones hacia las categorías mayoritarias, y la naturaleza lineal del modelo, que no logró capturar completamente las relaciones complejas entre las variables predictoras y la variable objetivo. Como resultado, el modelo fue más preciso en las clases mayoritarias y tuvo un desempeño muy pobre en las clases menos representadas.

b) Árboles de decisión

El modelo de árboles de decisión se implementó utilizando un proceso de validación cruzada y ajuste de hiperparámetros, lo que permitió seleccionar una configuración óptima para maximizar el rendimiento del modelo. Los mejores parámetros encontrados fueron: criterion='entropy', max_depth=None, min_samples_leaf=4 y min_samples_split=2. Esta combinación permitió al modelo capturar relaciones complejas sin imponer límites a la profundidad del árbol, al tiempo que se evitó el sobreajuste mediante restricciones mínimas en el número de muestras por hoja y por división.

Los árboles de decisión presentan menos restricciones que otros modelos estadísticos, lo que los hace adecuados para problemas con relaciones complejas entre variables. En este estudio, el conjunto de datos cumple con los supuestos esenciales para la aplicación del modelo. En primer lugar, no se requiere linealidad entre las variables predictoras y la variable dependiente, lo que permite detectar patrones no lineales de manera eficiente. Además, se verificó la independencia de las observaciones, asegurando que cada registro representa una entidad única sin dependencia estructural con otros datos. El modelo es robusto frente a problemas de multicolinealidad, permitiendo la inclusión de múltiples variables sin comprometer la estabilidad del algoritmo. El tamaño muestral es adecuado para evitar sobreajuste y garantizar una correcta generalización de los resultados. Finalmente, se aplicó SMOTE para mitigar el desbalance de clases, mejorando la representación de categorías con menor frecuencia en los datos.

El modelo alcanzó una exactitud global del 74%, lo que representa una mejora significativa respecto al modelo de regresión logística. Este valor indica que el modelo fue capaz de identificar patrones de manera más efectiva, incluso en clases que anteriormente eran difíciles de clasificar. A diferencia de versiones anteriores, donde las clases minoritarias no eran correctamente identificadas, esta nueva configuración mostró un mejor rendimiento en todas las clases.

La matriz de confusión mostrada en la Tabla 18, evidencia un desempeño sólido en todas las categorías. La Clase 5 (Solvencia Muy Bajo) fue la mejor clasificada, con 249 de las 250 observaciones correctamente predichas. Asimismo, la Clase 4 (Solvencia Bajo) obtuvo un alto desempeño, con 221 predicciones correctas sobre 252 observaciones. Las clases 1, 2 y 3 también fueron clasificadas con precisión moderada, aunque se observó cierta confusión entre las Clases 2 y 3, lo cual es esperable dado su comportamiento similar.

Tabla 18.
Matriz de confusión Árboles de decisión

	1	**2**	**3**	**4**	**5**
1	175	60	28	6	0
2	64	127	37	13	0
3	27	22	160	40	0
4	2	10	15	221	4
5	0	0	0	1	249

El reporte de clasificación mostrado en la Tabla 19, detalla las métricas clave del modelo como precisión, recall y F1-score. La Clase 5 (Solvencia Muy Bajo) fue la mejor clasificada, con un F1-score de 0.99, lo que refleja una excelente capacidad del modelo para identificar esta categoría. La Clase 4 (Solvencia Bajo) también mostró un desempeño destacado, con un F1-score de 0.83. Las clases 1 y 3 obtuvieron un F1-score de 0.65, mientras que la Clase 2 presentó el desempeño más bajo con un F1-score de 0.55, lo que sugiere que el modelo tuvo ciertas dificultades para distinguir esta categoría de otras cercanas.

El modelo obtuvo un Macro Promedio F1-Score de 74%, lo que refleja un desempeño equilibrado en la clasificación de todas las categorías, incluidas aquellas que anteriormente eran difíciles de predecir. En contraste con versiones anteriores del modelo, donde el rendimiento en las clases minoritarias era limitado, esta mejora indica que el modelo logró una mejor generalización y fue capaz de capturar patrones más diversos en los datos. Además, el Weighted Promedio F1-Score también fue de 74%, lo que demuestra que el modelo no solo mantuvo un

buen desempeño promedio, sino que también logró distribuir de manera efectiva su capacidad predictiva en función de la frecuencia de cada clase, reduciendo el sesgo hacia las categorías más representadas.

Tabla 19.
Reporte de clasificación Árboles de decisión

	PRECISIÓN	RECALL	F1-SCORE	SUPPORT
1	0.65	0.65	0.65	269
2	0.58	0.53	0.55	241
3	0.67	0.64	0.65	249
4	0.79	0.88	0.83	252
5	0.98	1	0.99	250
ACURRACY			0.74	1261
MACRO AVG	0.73	0.74	0.74	1261
WEIGHTED AVG	0.73	0.74	0.74	1261

Este modelo mejoró el desempeño global en comparación con la regresión logística, mostrando una mayor exactitud y un mejor equilibrio en la clasificación de todas las categorías, incluidas las menos representadas. Su capacidad para capturar relaciones no lineales lo hace especialmente efectivo en problemas complejos con estructuras de datos heterogéneas.

A pesar de estos desafíos, los resultados indican que los árboles de decisión son una herramienta útil para establecer una base sólida en el análisis predictivo, especialmente para identificar patrones en las categorías más frecuentes.

c) Random Forest

El modelo Random Forest se implementó eligiendo un ensamble de 25 árboles en un primer paso y, posteriormente, se utilizó una búsqueda de hiperparámetros para optimizar su desempeño. Utilizando GridSearchCV, se lograron identificar los mejores parámetros para el modelo, lo que permitió mejorar su rendimiento global. Los mejores hiperparámetros encontrados fueron 'max_depth': None,

'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200.

El modelo de Random Forest cumple con los criterios esenciales para su correcta implementación en este estudio. A diferencia de la regresión logística, no requiere que las variables predictoras tengan una relación lineal con la variable dependiente, lo que permite capturar relaciones complejas dentro de los datos. Se verificó la independencia de las observaciones, evitando dependencias que puedan influir en los resultados del modelo. Además, la combinación de múltiples árboles dentro del bosque ayuda a reducir problemas de sobreajuste, mejorando la capacidad de generalización del modelo. La selección de hiperparámetros mediante GridSearchCV aseguró un balance adecuado entre precisión y estabilidad, y la aplicación de SMOTE contribuyó a mejorar la distribución de clases en el conjunto de entrenamiento, reduciendo el sesgo hacia las categorías mayoritarias y permitiendo una mejor identificación de las cajas con menor solvencia

Una vez que los mejores parámetros fueron determinados, la evaluación del modelo alcanzó una exactitud global del 75%, lo que representa una mejora significativa y considerada respecto a los otros modelos evaluados. El desempeño logrado refleja la capacidad de Random Forest para manejar relaciones no lineales e interacciones complejas entre las variables predictoras, esto demuestra que este modelo es más efectivo en comparación con modelos más simples como la regresión logística, utilizada en el apartado anterior.

En el modelo implementado, se evalúa a La matriz de confusión, la que se muestra en la Tabla 20, podemos apreciar cómo el modelo clasificó las diferentes clases de solvencia. En la Clase 1 (Solvencia Muy Alto), el modelo logró predecir correctamente 178 de 269 observaciones, aunque debemos indicar que hubo algunas confusiones con la Clase 2 (Solvencia Alto). En esta Clase 2, el modelo alcanzó un desempeño sólido, con 131 de 241 observaciones correctamente clasificadas, logrando un recall del 54%. Sin embargo, la Clase 3 (Solvencia Moderado) mostró resultados más débiles, con solo 166 de 249 observaciones correctamente clasificadas, y muchas observaciones fueron confundidas con las clases adyacentes (Clases 1 y 2). En las Clases 4 y 5 (Solvencia Bajo y Muy Bajo),

el modelo no pudo predecir ninguna observación correctamente en la Clase 5, aunque en la Clase 4, logró un desempeño razonable con 221 observaciones de las 249 disponibles.

Tabla 20.

Matriz de confusión Random Forest

	1	2	3	4	5
1	178	57	28	6	0
2	64	131	33	13	0
3	22	21	166	40	0
4	2	10	15	221	4
5	0	0	0	0	250

El reporte de clasificación mostrado en la Tabla 21, detalla el F1-Score, que mide el equilibrio entre precisión y recall para cada clase. En general, el modelo tuvo el mejor desempeño en las Clases 4 (Solvencia Alto) y 5 (Solvencia Muy Alto), con un F1-Score de 0.83 para la Clase 4 y un F1-Score de 0.99 para la Clase 5, lo que indica que el modelo fue muy eficaz en identificar estas clases. Por otro lado, las Clases 1 y 2, con una menor representación, mostraron un F1-Score moderado de 0.67 y 0.57, respectivamente, mientras que la Clase 3 (Solvencia Moderado) tuvo el peor desempeño, con un F1-Score de 0.68.

Tabla 21.

Reporte de clasificación Random Forest

	PRECISIÓN	RECALL	F1-SCORE	SUPPORT
1	0.67	0.66	0.67	269
2	0.60	0.54	0.57	241
3	0.69	0.67	0.68	249
4	0.79	0.88	0.83	252
5	0.98	1.00	0.99	250
ACURRACY			0.75	1261
MACRO AVG	0.75	0.75	0.75	1261
WEIGHTED AVG	0.75	0.75	0.75	1261

También se puede observar que a nivel global, el Macro Promedio F1-Score fue de 0.75, lo que refleja un desempeño consistente en todas las clases, aunque con un sesgo hacia las clases mayoritarias. El Weighted Promedio F1-Score también fue de 0.75, lo que refuerza la idea de que el modelo tiene un desempeño más fuerte en las clases predominantes.

El modelo de Random Forest ha demostrado ser una mejora significativa respecto a los modelos anteriores, mostrando un buen desempeño global, especialmente en la clasificación de las categorías predominantes (Clase 2 y Clase 5). La capacidad de Random Forest para manejar relaciones no lineales y la complejidad de las interacciones entre variables lo convierte en una herramienta más robusta para problemas de clasificación como este.

d) Máquinas de Soporte Vectorial (SVM)

Finalmente se aplicó el modelo de Máquinas de Soporte Vectorial. Utilizando GridSearchCV, se optimizó el modelo, encontrando los mejores parámetros: C = 10, gamma = 'auto', kernel = 'rbf', y class_weight = None. Estos parámetros mejoraron el rendimiento del modelo, permitiéndole adaptarse mejor a las características del conjunto de datos.

El modelo de SVM cumple con los criterios esenciales para su correcta implementación. En primer lugar, se verificó la separabilidad de clases en el espacio de características, lo que permitió la construcción de un hiperplano óptimo para la clasificación de las observaciones. La independencia de las observaciones se mantuvo a lo largo del análisis, evitando sesgos en la predicción. Se aplicó StandardScaler para la normalización de los datos, asegurando que todas las variables tengan un impacto equitativo en la identificación del margen óptimo del clasificador. Además, se optimizaron parámetros del kernel mediante GridSearchCV, garantizando que el modelo capture adecuadamente las relaciones no lineales presentes en los datos. Por último, se utilizó SMOTE para mejorar el balance de clases y reducir el sesgo hacia las categorías mayoritarias, optimizando la capacidad de predicción en clases menos representadas.

El modelo alcanzó una exactitud global del 71%, lo que indica un buen desempeño, especialmente en la clasificación de las categorías mayoritarias. Sin embargo, a pesar de esta mejora en la exactitud global, el modelo aún muestra limitaciones al clasificar correctamente las clases menos representadas. Aunque la exactitud es relativamente alta, el modelo no aprovechó completamente sus capacidades para capturar relaciones complejas debido al desbalance de clases en el conjunto de datos.

La matriz de confusión que se puede observar en la Tabla 22, refleja cómo el modelo clasificó las diferentes clases de solvencia. En la Clase 1 (Solvencia Muy Alto), el modelo logró clasificar correctamente 122 de 192 observaciones, aunque una parte significativa fue confundida con la Clase 2 (Solvencia Alto). En la Clase 2 (Solvencia Alto), el modelo logró su mejor desempeño, con 177 predicciones correctas de 262 observaciones, obteniendo un recall del 68%. Sin embargo, en la Clase 3 (Solvencia Moderada), solo 13 de 65 observaciones fueron clasificadas correctamente, y muchas fueron confundidas con las Clases 1 y 2. Finalmente, en las Clases 4 y 5 (Solvencia Baja y Muy Baja), el modelo no logró predecir correctamente ninguna observación, clasificando esas instancias erróneamente como otras categorías. Esto pone en evidencia la debilidad del modelo en manejar datos desbalanceados.

Tabla 22.
Matriz de confusión SVM

	1	**2**	**3**	**4**	**5**
1	164	64	30	11	0
2	62	125	35	19	0
3	21	22	155	47	4
4	0	11	31	205	5
5	1	0	0	0	249

El reporte de clasificación (Tabla 23) muestra las métricas clave de precisión, recall y F1-score para cada clase. En términos de rendimiento, la Clase 2 (Solvencia Alto) fue la mejor clasificada, con un F1-score del 64%, seguida de la Clase 1 (Solvencia

Muy Alto) con un F1-score del 59%, lo que indica un desempeño razonablemente bueno en estas clases. Sin embargo, la Clase 3 (Solvencia Moderado) mostró un F1-score del 31%, con un recall bajo del 20%, lo que refleja la dificultad del modelo para generalizar y predecir correctamente esta categoría. Las Clases 4 y 5 (Solvencia Bajo y Muy Bajo) obtuvieron métricas nulas en precisión y recall, lo que resalta la incapacidad del modelo para identificar estas clases minoritarias.

A nivel general, podemos indicar que el modelo alcanzó un Macro Promedio F1-Score del 31%, lo que indica un desempeño débil en las categorías menos representadas. El Weighted Promedio F1-Score fue del 57%, reflejando un sesgo hacia las clases mayoritarias, lo que demuestra que el modelo es más preciso en las categorías más frecuentes.

Tabla 23.
Reporte de clasificación SVM

	PRECISIÓN	RECALL	F1-SCORE	SUPPORT
1	0.66	0.61	0.63	269
2	0.56	0.52	0.54	241
3	0.62	0.62	0.62	249
4	0.73	0.81	0.77	252
5	0.97	1.00	0.98	250
ACURRACY			0.71	1261
MACRO AVG	0.71	0.71	0.71	1261
WEIGHTED AVG	0.71	0.71	0.71	1261

El modelo de SVM mostró un desempeño aceptable en la clasificación de las categorías mayoritarias (clases 1 y 2), destacando su capacidad para capturar patrones en estas observaciones. Sin embargo, su desempeño en las categorías minoritarias (clases 4 y 5) fue deficiente, lo que pone de manifiesto una limitación en el manejo del desbalance de clases y la naturaleza compleja del problema.

e) Comparación de modelos

Se evaluaron cuatro algoritmos: Random Forest, Árboles de Decisión, SVM y Regresión Logística, con base en tres métricas principales:

- Exactitud (Accuracy): porcentaje de predicciones correctas sobre el total.
- F1-score macro promedio: mide el rendimiento considerando cada clase por igual, sin importar cuántos casos tenga cada una.
- F1-score ponderado (weighted): considera el tamaño de cada clase, útil cuando hay clases desbalanceadas.

Tabla 24.
Comparación de modelos evaluados

MODELO	EXACTITUD	MARCO PROMEDIO F1 - SCORE	WEIGHTED PROMEDIO F1 - SCORE
RANDOM FOREST	75%	0.75	0.75
ÁRBOLES DE DECISIÓN	74%	0.74	0.74
SVM	71%	0.71	0.71
REGRESIÓN LOGÍSTICA	61%	0.60	0.60

- Random Forest: Fue el mejor modelo. Obtuvo la mayor exactitud (75%) y valores consistentes tanto en el F1 macro como en el ponderado (0.75), lo que indica un buen rendimiento general en todas las clases. Es una opción adecuada para problemas multiclase con posibles desbalances.
- Árboles de Decisión: Mostró un rendimiento muy similar al de Random Forest, con una exactitud de 74% y valores de F1 macro y ponderado de 0.74. Aunque es ligeramente inferior, su desempeño sigue siendo competitivo.
- SVM: Alcanzó una exactitud de 71% y valores de F1 macro y ponderado de 0.71. Si bien el rendimiento es aceptable, es inferior a los modelos basados en árboles.
- Regresión Logística: Obtuvo la menor exactitud (61%) y también el menor F1-score (macro y ponderado de 0.60), lo que indica un desempeño más limitado para este tipo de problema

De acuerdo con los resultados observados en la Tabla 24, el modelo Random Forest fue el más robusto y equilibrado para predecir la solvencia, ya que logró alta precisión general y buen rendimiento en todas las clases, incluso las menos representadas. Por ello, se considera el modelo óptimo para esta investigación.

5.2. Discusión de resultados

5.2.1 Discusión sobre el análisis exploratorio de los datos.

El análisis exploratorio permitió identificar patrones relevantes en las variables macroeconómicas y microfinancieras, destacando que la mayoría de las cajas se encuentran en un nivel de solvencia Alto o Muy Alto. Este resultado refleja una gestión financiera razonable en términos generales, aunque ciertos indicadores como la rentabilidad y la eficiencia muestran brechas significativas.

- Solvencia predominante: El 85.24%de las cajas pertenecen a las categorías de solvencia “Muy Alto” o “Alto”. Este resultado subraya la fortaleza general del sistema de microfinanzas en el Perú, aunque es importante notar que las instituciones en las categorías “Baja” y “Muy Baja” representan riesgos críticos que podrían tener un impacto desproporcionado en el sistema financiero si no se manejan adecuadamente, por lo que requieren la debida atención y monitoreo.
- Impacto de la eficiencia y rentabilidad: Las bajas rentabilidades observadas (56.22% de las cajas presentan rentabilidad ”Bajo”) y las altas ineficiencias operativas evidenciadas en varias cajas sugieren que existen oportunidades claras para mejorar la sostenibilidad financiera.

Estos resultados iniciales son consistentes con la literatura previa sobre microfinanzas, que enfatiza que las instituciones de este tipo enfrentan retos relacionados con la gestión de riesgos y la eficiencia operativa.

Los resultados obtenidos en esta investigación coinciden con estudios previos, como el de Alison Jara en Ecuador (Jara Velastegui, 2023), que evidencia la importancia de indicadores financieros clave como liquidez, rentabilidad y eficiencia

en la evaluación de la salud financiera de las cooperativas de ahorro y crédito. De manera similar, los hallazgos del análisis Z-Score de Altman confirman las dificultades enfrentadas por muchas instituciones financieras en la región, alineándose con las tendencias observadas en las cajas municipales y rurales del Perú.

5.2.2 Discusión sobre la selección de variables.

Sobre el análisis hecho en el acápite de selección de variables, el análisis de correlación evidencia la importancia de las variables microfinancieras (rentabilidad, calidad de activos, eficiencia, liquidez) y macroeconómicas (inflación) en la predicción de la solvencia.

- Rentabilidad y eficiencia: Su impacto en la solvencia destaca la necesidad de gestionar adecuadamente los márgenes operativos y los costos. Estos resultados coinciden con estudios previos que señalan que la rentabilidad es un indicador clave de sostenibilidad financiera en instituciones de microfinanzas.
- Calidad de activos y liquidez: Subrayan la importancia de una cartera bien gestionada y una posición de liquidez sólida para garantizar la estabilidad financiera.
- Inflación: Aunque su impacto fue moderado (r = 0.164), refleja que los entornos macroeconómicos estables contribuyen a la solvencia financiera de las instituciones.

Estos hallazgos son consistentes con los resultados del estudio realizado por Yuly Franco en Colombia (Franco, 2024), que destaca que la integración de variables financieras y no financieras mejora significativamente la capacidad predictiva de los modelos de Machine Learning.

5.2.3 Discusión sobre los modelos de predicción.

En esta investigación se evaluó el impacto de diferentes modelos de Machine Learning para predecir la solvencia de las cajas rurales y municipales en el Perú.

Los resultados obtenidos ofrecen evidencia clave sobre el rendimiento de cada modelo y su capacidad para abordar un problema crítico en el sector financiero.

Los resultados son coherentes con estudios como el de Anastasios Petropoulos en Estados Unidos (Anastasios Petropoulos, 2020), que también encontró que Random Forest superó a otros métodos en la predicción de insolvencia bancaria, resaltando su capacidad para manejar relaciones no lineales y datos desbalanceados.

En comparación, el modelo de Árboles de Decisión obtuvo una exactitud ligeramente inferior del 74%, con un F1-score macro de 0.74 y un F1-score ponderado también de 0.74. Si bien su desempeño fue relativamente parejo, es ligeramente inferior al de Random Forest. A diferencia de lo observado en versiones simplificadas de este modelo, en este caso no se detectó un sesgo marcado hacia clases mayoritarias, aunque persiste una ligera disminución en el rendimiento para clases minoritarias como la Clase 5 (Solvencia Muy Baja).

SVM, por su parte, logró una exactitud del 71%, con un F1-score macro y ponderado de 0.71. Aunque su desempeño fue aceptable en clases frecuentes, el modelo mostró limitaciones en su capacidad de generalización hacia clases menos representadas, lo que es consistente con su naturaleza menos robusta frente a datos desbalanceados. Estos resultados coinciden con los hallazgos del estudio argentino de Sattler, Terreno y Pérez (2024), quienes también reportaron un mejor rendimiento de modelos no lineales como Random Forest frente a SVM y métodos lineales.

Finalmente, Regresión Logística, que sirvió como modelo base, tuvo el desempeño más bajo, con una exactitud del 61% y un F1-Score Macro Promedio de 0.60. Este modelo lineal, a pesar de su simplicidad y utilidad como punto de partida, presentó limitaciones inherentes al no poder capturar las relaciones no lineales en los datos. Además, mostró un sesgo hacia las clases mayoritarias, con un rendimiento muy bajo en las clases minoritarias, lo que afectó su capacidad de generalización.

Del análisis realizado, podemos concluir que Random Forest es el modelo más adecuado para este problema, ya que proporciona un equilibrio entre precisión, recall y capacidad de manejar las relaciones no lineales. Árboles de Decisión

también presentan una opción sólida, pero requieren mejoras en la clasificación de clases minoritarias. SVM y Regresión Logística, aunque útiles en algunos contextos, no son tan eficaces en este escenario debido a su incapacidad para abordar adecuadamente el desbalance de clases y las complejidades del problema de solvencia.

Los resultados de esta investigación confirman que el uso de técnicas avanzadas, como las propuestas por Nick Valdivia y Jean Rojas en Perú (Valdivia Saavedra & Rojas Munares, 2023), pueden mejorar significativamente la predicción de insolvencias en las cajas municipales y rurales. Mientras que las metodologías tradicionales tienen un valor histórico, los avances en Machine Learning ofrecen una precisión y versatilidad mayores, permitiendo abordar los desafíos dinámicos del sector financiero peruano.

5.3 Conclusión general

La presente investigación ha permitido identificar y evaluar los modelos de Machine Learning más efectivos para predecir la solvencia de las cajas rurales y municipales en el Perú, abordando un problema crítico para el fortalecimiento del sistema de microfinanzas. A partir del análisis comparativo de cuatro modelos (Regresión Logística, Árboles de Decisión, Random Forest y SVM), se concluyó que el modelo Random Forest es el más robusto y preciso para esta tarea, alcanzando una exactitud del 75% y destacando en las métricas de F1-Score (Macro y Weighted). Este modelo mostró una capacidad notable para manejar las complejidades propias de los datos financieros, incluyendo interacciones no lineales y un marcado desbalance entre clases.

La hipótesis inicial, que planteaba que los modelos más avanzados, como Random Forest, serían más eficaces debido a su habilidad para capturar patrones complejos, fue plenamente confirmada. Los resultados obtenidos evidenciaron que este modelo supera consistentemente a otros enfoques, como Árboles de Decisión y SVM, en precisión y generalización, mientras que la Regresión Logística mostró limitaciones significativas debido a su naturaleza lineal y su incapacidad para capturar las complejidades inherentes al problema.

Además, los hallazgos resaltan el potencial práctico de los modelos de Machine Learning para abordar desafíos clave en el sector financiero peruano, especialmente en la mejora de la gestión de riesgos y la toma de decisiones estratégicas. La implementación de un modelo como Random Forest podría facilitar la detección temprana de riesgos de insolvencia, optimizando recursos y promoviendo la sostenibilidad del sistema de cajas rurales y municipales. Este enfoque no solo contribuiría a la estabilidad financiera de las instituciones, sino que también reforzaría la confianza en el sector de microfinanzas como motor de desarrollo económico en el país.

Conclusiones

El estudio permitió concluir que el modelo Random Forest es el más robusto para predecir insolvencia, debido a su capacidad para manejar datos financieros complejos y desbalanceados. Este modelo alcanzó un 75% de exactitud y F1-Scores consistentes, superando a otros métodos evaluados. Su aplicación en la gestión de riesgos puede optimizar la asignación de recursos y mejorar la toma de decisiones estratégicas en las cajas municipales y rurales.

El análisis de correlación identificó como variables clave: rentabilidad, eficiencia, calidad de activos y liquidez, dentro de las variables microfinancieras, mientras que, en el ámbito macroeconómico, la inflación tuvo un impacto moderado. La rentabilidad y la eficiencia destacaron por su relación directa con la sostenibilidad financiera, mientras que la calidad de activos y la liquidez subrayaron la importancia de una gestión prudente de la cartera crediticia y los recursos disponibles. Estas conclusiones coinciden con estudios previos y refuerzan la relevancia de un enfoque integral que combine factores financieros y no financieros.

Random Forest demostró ser el modelo con mejor rendimiento, al combinar precisión y capacidad de generalización. Su ventaja radica en su arquitectura basada en múltiples árboles de decisión, lo que le permite manejar relaciones no lineales y preservar el rendimiento incluso ante desbalances de clases. Este comportamiento fue validado con métricas en validación cruzada y conjunto de prueba.

Los modelos de Machine Learning superaron consistentemente a los métodos tradicionales como la Regresión Logística, que obtuvo un rendimiento limitado (exactitud del 61%) y menor capacidad de detección de insolvencias. Esta diferencia resalta la necesidad de migrar hacia enfoques más sofisticados que permitan capturar relaciones complejas y mejorar la precisión de las predicciones.

Recomendaciones

Se recomienda incorporar el modelo Random Forest en la gestión de riesgos y evaluación crediticia de las cajas municipales y rurales del Perú, integrándolo a sistemas como Core Bancario, plataformas de Scoring y soluciones de Business Intelligence (Power BI, Tableau) mediante APIs o conectores. Esta integración garantizará su uso regular y automatizado en los procesos de análisis de solvencia.

Se sugiere establecer un proceso de validación mensual del modelo mediante el análisis de desempeño sobre nuevos datos reales, utilizando métricas como exactitud, F1-Score y matriz de confusión. Asimismo, es fundamental realizar una validación externa utilizando datos posteriores al periodo de entrenamiento, a fin de garantizar la capacidad predictiva del modelo en entornos cambiantes y reales.

Es necesario diseñar un plan de formación práctica dirigido al personal técnico y operativo de las cajas, enfocado en el uso y mantenimiento del modelo predictivo. Este plan debe incluir sesiones sobre interpretación de resultados, configuración del modelo, monitoreo de su rendimiento y actualización periódica. El objetivo es fortalecer la comprensión del modelo y fomentar una cultura de toma de decisiones basada en datos.

Se recomienda implementar un sistema de monitoreo continuo que permita el seguimiento de las variables más influyentes en la solvencia, tales como rentabilidad, eficiencia, calidad de activos, liquidez e inflación. Estas métricas deben integrarse en los reportes de gestión interna y formar parte de las reuniones estratégicas de evaluación de riesgos.

Se debe incentivar el uso de técnicas de Machine Learning, como Random Forest, frente a métodos tradicionales como la Regresión Logística, especialmente en contextos con estructuras de datos desbalanceadas o no lineales. Para ello, se recomienda organizar talleres y seminarios especializados en el sector financiero, orientados a difundir las ventajas técnicas y prácticas de estos modelos en la gestión moderna del riesgo crediticio.

Referencias

Alcalde, R., Alonso de Armiño, C., & García, S. (2022). Analysis of the economic sustainability of the supply chain sector by applying the Altman Z-score predictor. Sustainability, 14(2), 851. https://doi.org/10.3390/su14020851

Altman, E. (1968). Ratios financieros, análisis discriminante y predicción de quiebras corporativas. *The Journal of Finance*. doi:https://doi.org/10.2307/2978933

Altman, E., & Iwanicz-Drozdowska, M. (2016). Predicción de dificultades financieras en un contexto internacional: una revisión y análisis empírico del modelo Z-Score de Altman. *Revista de contabilidad y gestión financiera internacional*. doi:https://doi.org/10.1111/jifm.12053

Alvarez, M. (2023). *MicroCapital.* Retrieved from https://www.microcapital.org/news-wire-mexico-microfinance-industries-undergoing-major-changes/

Amat Rodrigo, J. (2017). Máquinas de Vector Soporte (Support Vector Machines, SVMs). Retrieved from https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

Anastasios Petropoulos, V. S. (2020). Predicción de insolvencias bancarias mediante técnicas de aprendizaje automático. *International Journal of Forecasting*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0169207019302663

Andrade, G., Herrera, D., & De Olloqui, F. (2015). *Inclusión financiera en América Latina y el Caribe: Coyuntura actual y desafíos para los próximos años.* doi:https://doi.org/10.18235/0000030

AWS. (2024). *¿Qué es la regresión logística?* Retrieved from https://aws.amazon.com/es/what-is/logistic-regression/

AWS. (2024). ¿Qué es la regresión logística? Retrieved from https://aws.amazon.com/es/what-is/logistic-regression/

AWS. (2024). *¿Qué es una red neuronal?* Retrieved from https://aws.amazon.com/es/what-is/neural-network/

Banco Santander. (2023). *Insolvencia: ¿qué es y cómo evitar este estado financiero?* Retrieved from https://www.santander.com/es/stories/insolvencia

BBVA. (2016). *¿Qué es la tasa de interés y por qué se cobra?* Retrieved from https://www.bbva.mx/educacion-financiera/creditos/ppi-que-es-la-tasa-de-interes.html

Chang Hidalgo, H. M., & Chirinos Mundaca, C. (2023). *Comparación de técnicas de estimación basadas en machine learning para predecir costos en los planes de adquisiciones de las entidades públicas del Perú.* Pimentel. Retrieved from https://repositorio.uss.edu.pe/bitstream/handle/20.500.12802/10566/Chang%20Hidalgo%2c%20Haward%20Miguel.pdf?sequence=1&isAllowed=y

Duana Dávila, D., Dueñas Soto, J., & Pérez Herrera, E. (2023). Retrieved from https://repository.uaeh.edu.mx/bitstream/bitstream/handle/123456789/20346/variables-macroeconomicas.pdf?sequence=1&isAllowed=y

Economista, E. (2022). *Inflación.* Retrieved from https://www.eleconomista.es/diccionario-de-economia/inflacion

Franco, Y. A. (2024). Machine Learning aplicado a dificultades financieras y quiebra empresarial: una revisión de literatura. *ResearchGate*. Retrieved from https://www.researchgate.net/profile/Nelson-Fonseca-Carreno/publication/371250033_libro_4_tomo_4_Ciencias_multidisciplinarias/links/647a24832cad460a1bee2f56/libro-4-tomo-4-Ciencias-multidisciplinarias.pdf#page=277

GanaMas. (2024). Cajas municipales disminuyeron sus utilidades en 77.8% a abril 2024. *GanaMas*. Retrieved from https://revistaganamas.com.pe/cajas-municipales-disminuyeron-sus-utilidades-en-77-8-a-abril-2024/

Hernandez, R. (2014). *Metodología de la investigación.* Retrieved from https://www.esup.edu.pe/wp-content/uploads/2020/12/2.%20Hernandez,%20Fernandez%20y%20Baptista-Metodolog%C3%ADa%20Investigacion%20Cientifica%206ta%20ed.pdf

IBM. (2021). *¿Qué es el random forest?* Retrieved from https://www.ibm.com/mx-es/topics/random-forest

IBM. (2021). *¿Qué es un árbol de decisión?* Retrieved from https://www.ibm.com/es-es/topics/decision-trees

IBM. (2021). *Funcionamiento de SVM.* Retrieved from https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works

IBM. (2022). *¿Qué es una red neuronal?* Retrieved from https://www.ibm.com/mx-es/topics/neural-networks

IBM. (2024). *¿Qué es machine learning (ML)?* Retrieved from https://www.ibm.com/mx-es/topics/machine-learning

Jara Velastegui, A. (2023). *Análisis de probabilidad de quiebra de las cooperativas de ahorro y crédito, segmento 1 en la zona 3, período 2019- 2021.* Riobamba. Retrieved from https://repositorioslatinoamericanos.uchile.cl/handle/2250/8146265

Medina Huaman, R. (2024). *Factores determinantes de la sostenibilidad de las cajas municipales y rurales durante el período 2000-2019.* Lima. Retrieved from https://repositorio.esan.edu.pe/server/api/core/bitstreams/667182de-af3a-4c12-8fa4-e0a7fe404b7c/content

MEF. (2015). *Conoce los conceptos Basicos para comprender la economia del país.* Lima. Retrieved from https://www.mef.gob.pe/en/?id=61:conoce-los-conceptos-basicos-para-comprender-la-economia-del-pais&option=com_content&language=en-GB&view=article&lang=en-GB

Ohlson, J. (1980). Los ratios financieros y la predicción probabilística de la quiebra. *Journal of Accounting Research*. doi:https://doi.org/10.2307/2490395

Pedraza Nájar, X. L., & Pérez Juárez, J. (2021). *Medición del work engagement y su relación con la comunicación, liderazgo y TIC en una empresa editorial mexicana.* Colombia. doi:https://doi.org/10.15332/s2145-1389.2019.0001.02

PRODUCE. (2024). Micro y pequeñas empresas representan el 99.2% del tejido empresarial peruano. Retrieved from https://www.gob.pe/institucion/tuempresa/noticias/898863-micro-y-pequenas-empresas-representan-el-99-2-del-tejido-empresarial-peruano

PRODUCE. (2024). Mipymes generaron S/337 000 millones en ventas en el Perú durante el 2022. Retrieved from https://www.gob.pe/institucion/produce/noticias/897936-produce-mipymes-generaron-s-337-000-millones-en-ventas-en-el-peru-durante-el-2022

Rebajatuscuentas. (2023). *Diferencias entre caja municipal y rural.* Retrieved from https://rebajatuscuentas.com/pe/blog/diferencias-entre-caja-municipal-y-caja-rural

Rivera Martinez, I. (2019). *Comunicación interna y desempeño laboral de los empleados del hospital La Carlota en Montemorelos.* Mexico. doi:https://dspace.um.edu.mx/handle/20.500.11972/1047

Rojas Ramos , M., & Cruz Quispe, E. (2021). *Relación entre el Engagement y el Desempeño Laboral en la Empresa Andoriña Tours S.R.L.* Arequipa. Retrieved from https://repositorio.utp.edu.pe/bitstream/handle/20.500.12867/5460/M.Rojas_E.Cruz_Trabajo_de_Suficiencia_Profesional_Titulo_Profesional_2021.pdf?sequence=1&isAllowed=y

Rossi Valverde, R. (2022). Análisis del riesgo de quiebra de instituciones financieras peruanas, 2015-2021. *Revista Mexicana de Economía y Finanzas, Nueva Época*. Retrieved from https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462022000300101&lang=es

Sattler, S., Terreno, D., & Pérez, J. (2024). Un modelo jerárquico para la predicción de insolvencia empresarial. Aplicación de análisis discriminante y árboles de clasificación. *Cuadernos de economias ISSN 0121-4772*. doi:https://doi.org/10.15446/cuad.econ.v43n91.105115

SFC. (2021). *Actualidad de Sistema Financiero Colombiano.* Retrieved from https://fundpro.com/wdpr/wp-content/uploads/2022/03/comsectorfinanciero122021.pdf

Toledo, E. (2020). Microfinanzas en el Perú y los desafíos de la bancarización. *ResearchGate*. doi:http://dx.doi.org/10.21574/remipe.v4i1.111

Umair Ali, S. F. (2018). Enfoques de aprendizaje automático para predecir la quiebra de empresas: un estudio comparativo. doi:http://dx.doi.org/10.21203/rs.3.rs-4961599/v1

Valdivia Saavedra, N., & Rojas Munares, J. (2023). *Análisis de la situación financiera de las empresas del sector financiero del Perú aplicando el modelo Z de Altman, antes y después de la pandemia.* Lima. Retrieved from https://repositorioacademico.upc.edu.pe/handle/10757/669207

Valentina Lagasio, F. P. (2022). Evaluación de los determinantes del impago bancario mediante aprendizaje automático. *Information Sciences*. doi:https://doi.org/10.1016/j.ins.2022.10.128

Vázquez Burguillo, R. (2018). *Tasa de desempleo.* Retrieved from https://economipedia.com/definiciones/tasa-de-desempleo-paro.html

WORLD BANK GROUP. (2021). *Visualización del resumen ejecutivo.* Retrieved from https://www.worldbank.org/en/publication/globalfindex/interactive-executive-summary-visualization

Zhou, L., Keung Lai, K., & Yen, J. (2010). Predicción de quiebras incorporando variables macroeconómicas mediante redes neuronales. *ResearhGate*. doi:http://dx.doi.org/10.1109/TAAI.2010.24

Anexos

Código en Python

1. IMPORTAR LIBRERIAS

```
## ---- MANEJO DE DATOS Y GRAFICOS -------
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

## ---- ESTADISTICAS ------
import statsmodels.api as sm
from scipy import stats

## ---- REGRESIÓN LOGÍSTICA ------
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report, roc_auc_score, roc_curve
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
import matplotlib.pyplot as plt

## ---- ÁRBOLES DE DECISIÓN ------
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV, train_test_split
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
```

```
## ---- RANDOM FOREST ------
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE

## ---- SVM ------
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
```

1.1 MÉTODOS DE USO

```
[2] def Analisis_Count(DF,df):
      count = DF.value_counts()
      df_count = pd.DataFrame(count)
      df_count['%'] = (df_count/len(df)*100).round(2)
      return [count,df_count]
```

2. CARGAR DATASET

```
[3]  df = pd.read_csv("BaseDatos.csv", sep=";", encoding="latin1")
     df
```

3. ANÁLISIS EXPLORATORIO

3.1 ESTRUCTURA DEL DATASET

```
#Dimensionalidad
df.shape
```

```
[6] #Nombre de columnas
df.columns
```

```
#Tipo de datos
df.info()
```

```
df.describe()
```

3.2 VALORES MISSING

```
# Reemplazar los valores 'n.a.' por NaN para facilitar el cálculo si es necesario
df.replace('n.a.', np.nan, inplace=True)
```

```
[10] #Porcentaje de valores nulos por variable
     missing  = df.isnull()
     missingDf = pd.DataFrame(missing.sum(),columns=['COUNTS'])
     missingDf['%'] = (missing.sum()/len(df)*100).round(2)
     missingDf
```

```
#Visualización de valores nulos
plt.figure(figsize=(6,3))
sns.heatmap(missing,cbar=False)
plt.title(" DATOS MISSING")
```

```
# Eliminar filas con valores faltantes
df.dropna(inplace=True)
```

3.3 CATEGORIZACIÓN DE VARIABLES

```python
# Clasificación para PBI (%)
def PBI_CATEGORIA(PBI):
    if PBI > 7:
        return "1"  #1. Muy alto (> 7%)
    elif 5 < PBI <= 7:
        return "2"  #2. Alto (5% - 7%)
    elif 2 < PBI <= 5:
        return "3"  #3. Moderado (2% - 5%)
    elif 0 < PBI <= 2:
        return "4"  #4. Bajo (0% - 2%)
    else:
        return "5"  #5. Negativo (< 0%):

# Aplicar la función de clasificación a la columna PBI
df['PBI_CATEGORIA'] = df['PBI'].apply(PBI_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.PBI_CATEGORIA,df)[1]
```

```python
# Clasificación para INFLACION (%)
def INFLACION_CATEGORIA(INFLACION):
    if INFLACION < 1:
        return "1"  #1. Muy baja (< 1%)
    elif 1 <= INFLACION < 3:
        return "2"  #2. Baja (1% - 3%)
    elif 3 <= INFLACION < 6:
        return "3"  #3. Moderada (3% - 6%)
    elif 6 <= INFLACION < 10:
        return "4"  #4. Alta (6% - 10%)
    else:
        return "5"  #5. Muy alta (> 10%)

# Aplicar la función de clasificación a la columna INFLACION
df['INFLACION_CATEGORIA'] = df['INFLACION'].apply(INFLACION_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.INFLACION_CATEGORIA,df)[1]
```

```python
# Clasificación para DESEMPLEO (%)
def DESEMPLEO_CATEGORIA(desempleo):
    if desempleo < 3:
        return "1"  #1. Muy baja (< 3%)
    elif 3 <= desempleo < 5:
        return "2"  #2. Baja (3% - 5%)
    elif 5 <= desempleo < 10:
        return "3"  #3. Moderada (5% - 10%)
    elif 10 <= desempleo < 15:
        return "4"  #4. Alta (10% - 15%)
    else:  # desempleo >= 15%
        return "5"  #5. Muy alta (> 15%)

# Aplicar la función de clasificación a la columna DESEMPLEO
df['DESEMPLEO_CATEGORIA'] = df['DESEMPLEO'].apply(DESEMPLEO_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.DESEMPLEO_CATEGORIA,df)[1]
```

```python
# Clasificación para SOLVENCIA (%)
def SOLVENCIA_CATEGORIA(solvencia):
    if solvencia > 15:
        return "1"  # 1.  Muy alto (> 15%)
    elif 12 <= solvencia <= 15:
        return "2"  # 2.  Alto (12% - 15%)
    elif 10 <= solvencia < 12:
        return "3"  # 3.  Moderado (10% - 12%)
    elif 8 <= solvencia < 10:
        return "4"  # 4.  Bajo (8% - 10%)
    else:
        return "5"  # 5.  Muy bajo (< 8%)

# Aplicar la función de clasificación a la columna SOLVENCIA
df['SOLVENCIA_CATEGORIA'] = df['SOLVENCIA'].apply(SOLVENCIA_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.SOLVENCIA_CATEGORIA,df)[1]
```

```python
# Clasificación para CALIDAD ACTIVOS (%)
def CALIDAD_ACTIVOS_CATEGORIA(calidad_activos):
    if calidad_activos < 2:
        return "1"  #1. Muy bajo (< 2%)
    elif 2 <= calidad_activos < 5:
        return "2"  #2. Bajo (2% - 5%)
    elif 5 <= calidad_activos < 10:
        return "3"  #3. Moderado (5% - 10%)
    elif 10 <= calidad_activos < 15:
        return "4"  #4. Alto (10% - 15%)
    else:
        return "5"  #5. Muy alto (> 15%)

# Aplicar la función de clasificación a la columna CALIDAD ACTIVOS
df['CALIDAD_ACTIVOS_CATEGORIA'] = df['CALIDAD ACTIVOS'].apply(CALIDAD_ACTIVOS_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.CALIDAD_ACTIVOS_CATEGORIA,df)[1]
```

```python
# Clasificación para EFICIENCIA (%)
def EFICIENCIA_CATEGORIA(eficiencia):
    if eficiencia < 30:
        return "1"  #1. Muy bajo (< 30%)
    elif 30 <= eficiencia < 40:
        return "2"  #2. Bajo (30% - 40%)
    elif 40 <= eficiencia < 50:
        return "3"  #3. Moderado (40% - 50%)
    elif 50 <= eficiencia < 60:
        return "4"  #4. Alto (50% - 60%)
    else:
        return "5"  #5. Muy alto (> 60%)

# Aplicar la función de clasificación a la columna EFICIENCIA
df['EFICIENCIA_CATEGORIA'] = df['EFICIENCIA'].apply(EFICIENCIA_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.EFICIENCIA_CATEGORIA,df)[1]
```

```python
# Clasificación para RENTABILIDAD (%)
def RENTABILIDAD_CATEGORIA(rentabilidad):
    if rentabilidad > 20:
        return "1"  #1. Muy alto (> 20%)
    elif 15 <= rentabilidad <= 20:
        return "2"  #2. Alto (15% - 20%)
    elif 8 <= rentabilidad < 15:
        return "3"  #3. Moderado (8% - 15%)
    elif 5 <= rentabilidad < 8:
        return "4"  #4. Bajo (5% - 8%)
    else:
        return "5"  #5. Muy bajo (< 5%)

# Aplicar la función de clasificación a la columna RENTABILIDAD
df['RENTABILIDAD_CATEGORIA'] = df['RENTABILIDAD'].apply(RENTABILIDAD_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.RENTABILIDAD_CATEGORIA,df)[1]
```

```python
# Clasificación para LIQUIDEZ (%)
def LIQUIDEZ_CATEGORIA(liquidez):
    if liquidez > 30:
        return "1"  #1. Muy alto (> 30%)
    elif 25 <= liquidez <= 30:
        return "2"  #2. Alto (25% - 30%)
    elif 15 <= liquidez < 25:
        return "3"  #3. Moderado (15% - 25%)
    elif 10 <= liquidez < 15:
        return "4"  #4. Bajo (10% - 15%)
    else:
        return "5"  #5. Muy bajo (< 10%)

# Aplicar la función de clasificación a la columna LIQUIDEZ
df['LIQUIDEZ_CATEGORIA'] = df['LIQUIDEZ'].apply(LIQUIDEZ_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.LIQUIDEZ_CATEGORIA,df)[1]
```

```python
# Clasificación para POSICIÓN ME (%)
def POSICION_ME_CATEGORIA(posicion_me):
    if posicion_me < 5:
        return "1"  #1. Muy bajo (< 5%)
    elif 5 <= posicion_me < 10:
        return "2"  #2. Bajo (5% - 10%)
    elif 10 <= posicion_me < 20:
        return "3"  #3. Moderado (10% - 20%)
    elif 20 <= posicion_me < 30:
        return "4"  #4. Alto (20% - 30%)
    else:
        return "5"  #5. Muy alto (> 30%)

# Aplicar la función de clasificación a la columna POSICIÓN ME
df['POSICION_ME_CATEGORIA'] = df['POSICION ME'].apply(POSICION_ME_CATEGORIA)

# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.POSICION_ME_CATEGORIA,df)[1]
```

3.4 VARIABLE OBJETIVO

```
# Resumen general de la variable
print("Resumen de la variable 'SOLVENCIA_CATEGORIA':")
print(df['SOLVENCIA_CATEGORIA'].describe())
```

```
# Contar la cantidad y proporción de cada categoría
Analisis_Count(df.SOLVENCIA_CATEGORIA,df)[1]
```

```
# Gráfico de barras
plt.figure(figsize=(8, 6))
sns.countplot(x='SOLVENCIA_CATEGORIA', data=df, palette='viridis')
plt.title('Distribución de la variable SOLVENCIA_CATEGORIA', fontsize=14)
plt.xlabel('Categorías de Solvencia')
plt.ylabel('Frecuencia')
plt.xticks(rotation=0)
plt.show()
```

3.5 VARIABLES INDEPENDIENTES

```
#Variables numericas
Variables_numericas = ['ANO','MES','PBI_CATEGORIA', 'INFLACION_CATEGORIA', 'DESEMPLEO_CATEGORIA',
                       'SOLVENCIA_CATEGORIA', 'CALIDAD_ACTIVOS_CATEGORIA',
                       'EFICIENCIA_CATEGORIA', 'RENTABILIDAD_CATEGORIA', 'LIQUIDEZ_CATEGORIA',
                       'POSICION_ME_CATEGORIA']

# Crear un DataFrame solo con las variables numéricas seleccionadas
df_numericas = df[Variables_numericas]
```

```
[27] #Tipo de datos
     df.info()
```

3.5.1 GRÁFICO DE BARRAS

```
# Gráfico de barras para frecuencia
# Configuración de filas y columnas para los subgráficos
n_cols = 3  # Número de gráficos por fila
n_rows = int(np.ceil(len(df_numericas.columns) / n_cols))  # Filas necesarias

# Crear el gráfico
plt.figure(figsize=(n_cols * 5, n_rows * 5))  # Ajustar el tamaño de la figura

for i, column in enumerate(df_numericas.columns, 1):
    plt.subplot(n_rows, n_cols, i)  # Posicionar subgráfico
    df_numericas[column].value_counts().sort_index().plot(kind='bar', color='skyblue', edgecolor='black')
    plt.title(f'Frecuencia de {column}', fontsize=12)  # Título del subgráfico
    plt.xlabel(column, fontsize=10)  # Etiqueta del eje X
    plt.ylabel('Frecuencia', fontsize=10)  # Etiqueta del eje Y
    plt.grid(axis='y', linestyle='--', alpha=0.7)  # Agregar una cuadrícula en el eje Y

plt.tight_layout()  # Ajustar diseño
plt.show()
```

3.5.2 HISTOGRAMAS

```
# Graficar histogramas para todas las variables numéricas
df_numericas.hist(bins=50, figsize=(20, 15))
plt.suptitle('Histograma de Variables Numéricas', fontsize=16)
plt.show()
```

3.5.3 BOXPLOT

```
# Calcular el número de filas y columnas necesarias para los subgráficos
n_vars = len(df_numericas.columns)
n_cols = 3  # Número de columnas (puedes ajustarlo según lo necesites)
n_rows = (n_vars // n_cols) + (n_vars % n_cols > 0)  # Número de filas necesarias

# Graficar boxplot para cada variable numérica
plt.figure(figsize=(n_cols*5, n_rows*5))
for i, column in enumerate(df_numericas.columns, 1):
    plt.subplot(n_rows, n_cols, i)
    sns.boxplot(data=df, x=column)
    plt.title(f'Boxplot de {column}')

plt.tight_layout()
plt.show()
```

3.6 CORRELACIÓN DE VARIABLES

```python
# Calcular la matriz de correlación
correlation_matrix = df_numericas.corr()

# Graficar el heatmap de la matriz de correlación
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('Matriz de Correlación')
plt.show()
```

```python
corr = df_numericas.corr()
corr[['SOLVENCIA_CATEGORIA']].sort_values(by = 'SOLVENCIA_CATEGORIA',ascending = False)
```

4. MODELOS DE PREDICCION

4.1 REGRESIÓN LOGÍSTICA

```
[33] #Definir las variables predictoras (X) y la variable objetivo (y)
     X = df[['RENTABILIDAD_CATEGORIA', 'CALIDAD_ACTIVOS_CATEGORIA',
             'LIQUIDEZ_CATEGORIA', 'INFLACION_CATEGORIA', 'EFICIENCIA_CATEGORIA']]
     y = df['SOLVENCIA_CATEGORIA']

     #Escalado de variables predictoras
     scaler = StandardScaler()
     X = scaler.fit_transform(X)

     #Manejar el desbalanceo de clases
     smote = SMOTE(random_state=42)
     X, y = smote.fit_resample(X, y)

     #Dividir los datos en conjuntos de entrenamiento y prueba
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

     #Ajuste de hiperparámetros usando GridSearchCV
     param_grid = {
         'C': [0.001, 0.01, 0.1, 1, 10, 100],
         'penalty': ['l1', 'l2'],
         'solver': ['liblinear', 'saga']
     }
```

```
     }

     logreg_model = LogisticRegression(max_iter=1000, random_state=42)
     grid_search = GridSearchCV(logreg_model, param_grid, cv=5, scoring='accuracy')
     grid_search.fit(X_train, y_train)

     #Obtener el mejor modelo
     best_logreg_model = grid_search.best_estimator_
     print(f"Mejores hiperparámetros: {grid_search.best_params_}")

     #Realizar predicciones con el mejor modelo
     y_pred = best_logreg_model.predict(X_test)
     y_pred_prob = best_logreg_model.predict_proba(X_test)[:, 1]

     #Evaluar el modelo
     #Exactitud
     accuracy = accuracy_score(y_test, y_pred)
     print(f'\nExactitud del modelo: {accuracy:.2f}')

     #Matriz de confusión
     conf_matrix = confusion_matrix(y_test, y_pred)
     print("\nMatriz de Confusión:")
     print(conf_matrix)

     #Reporte de clasificación
     class_report = classification_report(y_test, y_pred)
     print("\nReporte de Clasificación:")
     print(class_report)
```

4.2 ÁRBOLES DE DECISIÓN

```
# 1. Definir las variables predictoras (X) y la variable objetivo (y)
X = df[['RENTABILIDAD_CATEGORIA', 'CALIDAD_ACTIVOS_CATEGORIA',
        'LIQUIDEZ_CATEGORIA', 'INFLACION_CATEGORIA', 'EFICIENCIA_CATEGORIA']]
y = df['SOLVENCIA_CATEGORIA']

# 2. Escalar las variables predictoras
scaler = StandardScaler()
X = scaler.fit_transform(X)

# Manejar el desbalanceo de clases
smote = SMOTE(random_state=42)
X, y = smote.fit_resample(X, y)

# 3. Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 4. Crear el modelo base de Árbol de Decisión
dt_model = DecisionTreeClassifier(random_state=42)

# 5. Definir los parámetros para GridSearchCV
param_grid = {
    'max_depth': [None, 2, 4, 6, 8, 10],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'criterion': ['gini', 'entropy']
}
```

```
# Configurar GridSearchCV
grid_search = GridSearchCV(estimator=dt_model, param_grid=param_grid, cv=5, scoring='accuracy', n_jobs=-1)

# 6. Ajustar el modelo con los datos de entrenamiento
grid_search.fit(X_train, y_train)

# 7. Obtener los mejores parámetros
best_params = grid_search.best_params_
print(f"Mejores parámetros encontrados: {best_params}")

# 8. Evaluar el modelo con los mejores parámetros
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test)

# 9. Métricas de evaluación
# Exactitud
accuracy = accuracy_score(y_test, y_pred)
print(f'\nExactitud del modelo: {accuracy:.2f}')

# Matriz de confusión
conf_matrix = confusion_matrix(y_test, y_pred)
print("\nMatriz de Confusión:")
print(conf_matrix)

# Reporte de clasificación
class_report = classification_report(y_test, y_pred)
print("\nReporte de Clasificación:")
print(class_report)
```

4.3 RANDOM FOREST

```python
# Definir las variables predictoras (X) y la variable objetivo (y)
X = df[['RENTABILIDAD_CATEGORIA','CALIDAD_ACTIVOS_CATEGORIA','LIQUIDEZ_CATEGORIA',
        'INFLACION_CATEGORIA','EFICIENCIA_CATEGORIA']]
y = df['SOLVENCIA_CATEGORIA']

# Preprocesamiento: Escalado y manejo del desbalanceo
scaler = StandardScaler()
X = scaler.fit_transform(X)
smote = SMOTE(random_state=42)
X, y = smote.fit_resample(X, y)

# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Definir la grilla de hiperparámetros para RandomForest
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2']
}


# Inicializar y entrenar el modelo Random Forest con GridSearchCV
rf_model = RandomForestClassifier(random_state=42)
grid_search = GridSearchCV(estimator=rf_model, param_grid=param_grid, cv=5, scoring='accuracy', n_jobs=-1)
grid_search.fit(X_train, y_train)
```

```python
# Obtener el mejor modelo y sus hiperparámetros
best_rf_model = grid_search.best_estimator_
print(f"Mejores hiperparámetros: {grid_search.best_params_}")

# Realizar predicciones en el conjunto de prueba
y_pred = best_rf_model.predict(X_test)

# Evaluar el modelo
accuracy = accuracy_score(y_test, y_pred)
print(f'Exactitud del modelo: {accuracy:.4f}')
print("\nMatriz de Confusión:")
print(confusion_matrix(y_test, y_pred))
print("\nReporte de clasificación:")
print(classification_report(y_test, y_pred))
```

4.4 MÁQUINAS DE SOPORTE VECTORIAL (SVM)

```
# Definir las variables predictoras (X) y la variable objetivo (y)
X = df[['RENTABILIDAD_CATEGORIA','CALIDAD_ACTIVOS_CATEGORIA','LIQUIDEZ_CATEGORIA',
        'INFLACION_CATEGORIA','EFICIENCIA_CATEGORIA']]
y = df['SOLVENCIA_CATEGORIA']

# Escalar las características
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Manejo de clases desbalanceadas
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_scaled, y)

# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.2, random_state=42)

# Definir los parámetros para GridSearchCV
param_grid = {
    'C': [0.1, 1, 10],
    'gamma': ['scale', 'auto'],
    'kernel': ['rbf', 'linear'],
    'class_weight': ['balanced', None]
}

# Usar GridSearchCV para encontrar los mejores hiperparámetros
grid_search = GridSearchCV(SVC(), param_grid, cv=5, scoring='accuracy', n_jobs=-1)
grid_search.fit(X_train, y_train)
```

```
# Obtener el mejor modelo
best_svm_model = grid_search.best_estimator_
print(f"Mejores Hiperparámetros: {grid_search.best_params_}")

# Realizar predicciones con el mejor modelo
y_pred = best_svm_model.predict(X_test)

# Evaluar el modelo
accuracy = accuracy_score(y_test, y_pred)
print(f'Exactitud del modelo: {accuracy:.2f}')

# Matriz de confusión
conf_matrix = confusion_matrix(y_test, y_pred)
print("\nMatriz de Confusión:")
print(conf_matrix)

# Reporte de clasificación
class_report = classification_report(y_test, y_pred)
print("\nReporte de clasificación:")
print(class_report)
```