



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado

Facultad de Ingeniería de Sistemas e Informática

Unidad de Posgrado

**Modelo de aprendizaje automático para mejorar la
predicción de la quiebra de empresas**

TESIS

Para optar el Grado Académico de Doctor en Ingeniería de

Sistemas e Informática

AUTOR

Dino Michael QUINTEROS NAVARRO

ASESOR

Dr. Ciro RODRIGUEZ RODRIGUEZ

Lima, Perú

2025



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

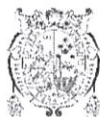
Referencia bibliográfica

Quinteros, D. (2025). *Modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas*. [Tesis de doctorado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Dino Michael Quinteros Navarro
Tipo de documento de identidad	DNI
Número de documento de identidad	41567782
URL de ORCID	https://orcid.org/0000-0001-8174-8771
Datos de asesor	
Nombres y apellidos	Ciro Rodriguez Rodriguez
Tipo de documento de identidad	DNI
Número de documento de identidad	06020241
URL de ORCID	https://orcid.org/0000-0003-2112-1349
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Cayo Víctor León Fernández
Tipo de documento	DNI
Número de documento de identidad	07001405
Miembro del jurado 1	
Nombres y apellidos	Frank Edmundo Escobedo Bailón
Tipo de documento	DNI
Número de documento de identidad	41671087
Miembro del jurado 2	
Nombres y apellidos	José Alfredo Herrera Quispe
Tipo de documento	DNI
Número de documento de identidad	40362859
Datos de investigación	
Línea de investigación	C.0.3.3. Desarrollo de modelos y aplicación de las tecnologías de información y comunicaciones.

Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Latitud: -12.045508 Longitud: -77.030957
Año o rango de años en que se realizó la investigación	2023 – 2024
URL de disciplinas OCDE	Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Vicedecanato de Investigación y Posgrado
Unidad de Posgrado

ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE DOCTOR EN INGENIERÍA DE SISTEMAS E INFORMÁTICA

A los veintiún (21) días del mes de febrero de 2025, siendo las 3:00 pm., se reunieron en el Auditorio, Profesor: Alfredo Celso Alva Bravo, el Jurado de Tesis conformado por los siguientes docentes:

Dr. Cayo Víctor León Fernández (Presidente)
Dr. Frank Edmundo Escobedo Bailón (Miembro)
Dr. José Alfredo Herrera Quispe (Miembro)
Dr. Ciro Rodríguez Rodríguez (Miembro Asesor)

Se inició la Sustentación invitando al candidato a Doctor **DINO MICHAEL QUINTEROS NAVARRO**, para que realice la exposición oral de la tesis para optar el Grado Académico de Doctor en Ingeniería de Sistemas e Informática, siendo la Tesis intitulada:

“MODELO DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA PREDICCIÓN DE LA QUIEBRA DE EMPRESAS”

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, habiendo obtenido la siguiente calificación:

DIECIOCHO (18) MUY BUENO

Por tanto, el presidente del Jurado, de acuerdo con el Reglamento General de Estudios de Posgrado, otorga al Magister **DINO MICHAEL QUINTEROS NAVARRO** el Grado Académico de Doctor en Ingeniería de Sistemas e Informática.

Siendo las 16:20 horas, el presidente del Jurado de Tesis, da por concluido el acto académico de Sustentación de Tesis.

Dr. Cayo Víctor León Fernández
(Presidente)

Dr. Frank Edmundo Escobedo Bailón
(Miembro)

Dr. José Alfredo Herrera Quispe
(Miembro)

Dr. Ciro Rodríguez Rodríguez
(Miembro Asesor)



UNIVERSIDAD NACIONAL MAYOR DE
SAN MARCOS
Universidad del Perú, Decana de América

CERTIFICADO DE SIMILITUD

Yo, **Ciro Rodriguez Rodriguez** en mi condición de asesor acreditado con **Dictamen N° 000113-2023-UPG-VDIP-FISI/UNMSM** de la tesis, cuyo título es “**Modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas**”, presentado por el magíster **Dino Michael Quinteros Navarro** para optar el grado de **Doctor en Ingeniería de Sistemas e Informática**, CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el 7% de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado/ título/ especialidad correspondiente.

Lima, 02 de diciembre de 2024

Firma del Asesor

Nombres y Apellidos del asesor: **Ciro Rodriguez Rodriguez**

DNI: **06020241**

Índice general

Índice general	ii
Lista de tablas.....	iv
Lista de figuras.....	vi
Dedicatoria	viii
Agradecimiento	ix
Resumen.....	x
Abstract	xi
I. INTRODUCCIÓN	1
1.1. Situación Problemática	1
1.2. Formulación del Problema.....	4
1.2.1. Problema General.....	4
1.2.2. Problemas Específicos	4
1.3. Justificación teórica	5
1.4. Justificación práctica.....	5
1.5. Objetivos.....	5
1.5.1. Objetivo general	5
1.5.2. Objetivos específicos	6
1.6. Hipótesis	6
1.6.1. Hipótesis general	6
1.6.2. Hipótesis específicas:.....	6
II. MARCO TEÓRICO	7
2.1. Marco filosófico o epistemológico de la investigación.....	7
2.2. Marco filosófico	7
2.3. Marco epistemológico	8
2.4. Antecedentes de investigación.....	9
2.5. Bases Teóricas.....	13
2.5.1. Quiebra empresarial.....	13
2.5.2. Rendimiento de pago	15
2.5.3. Riesgo de crédito	15
2.5.4. Inteligencia Artificial	16
2.5.5. Aprendizaje automático	17
2.5.6. Clasificación de los modelos de predicción	17
III. METODOLOGÍA.....	19
3.1. Tipo y diseño de investigación.....	19

3.2. Unidad de análisis	20
3.3. Población y muestra	20
3.4. Selección de la muestra	21
3.5. Recolección de datos	21
3.6. Confiabilidad de datos	22
3.7. Análisis e interpretación de datos	23
3.8. Dataset.....	25
3.9. Modelo de aprendizaje automático propuesto (MO2A)	26
3.9.1. Método.....	26
3.9.2. Modelo propuesto.....	26
3.9.3. Etapas del modelo propuesto	28
IV. RESULTADOS Y DISCUSIÓN	38
4.1. Análisis, interpretación y discusión de resultados	38
4.1.1. Análisis de resultados del preprocesamiento	38
4.1.2. Análisis de resultados de evaluación del performance de los algoritmos	39
4.1.3. Análisis de la correlación de factores relevantes de las empresas	50
4.1.4. Análisis de resultados de las métricas de performance del modelo optimizado	72
4.2. Pruebas de hipótesis.....	77
4.3. Presentación de resultados	93
V. CONCLUSIONES	99
VI. RECOMENDACIONES	101
REFERENCIAS BIBLIOGRÁFICAS	102
ANEXOS	107

Lista de tablas

<i>Tabla 1.</i> Diagrama del diseño de investigación	20
<i>Tabla 2.</i> Validez del instrumento	23
<i>Tabla 3.</i> Variables, tipos y definición	25
<i>Tabla 4.</i> Limpieza de datos	30
<i>Tabla 5.</i> Normalización Min Max.....	31
<i>Tabla 6.</i> Imputación KNN.....	31
<i>Tabla 7.</i> Outliers	32
<i>Tabla 8.</i> Compararon diversos algoritmos.....	35
<i>Tabla 9.</i> Matriz de línea base.....	38
<i>Tabla 10.</i> Matriz de datos preprocesado.....	38
<i>Tabla 11.</i> Resultados antes y después del preprocesamiento.....	39
<i>Tabla 12.</i> Matriz de confusión	40
<i>Tabla 13.</i> Resultados XGBoost	40
<i>Tabla 14.</i> Matriz de confusión	41
<i>Tabla 15.</i> Resultados SVM	42
<i>Tabla 16.</i> Matriz de confusión	43
<i>Tabla 17.</i> Resultados LSTM	43
<i>Tabla 18.</i> Matriz de confusión de regresión logística	45
<i>Tabla 19.</i> Resultados del modelo de regresión logística	45
<i>Tabla 20.</i> Matriz de confusión de bosques aleatorios.....	46
<i>Tabla 21.</i> Resultados del modelo de bosques aleatorios	47
<i>Tabla 22.</i> Matriz de confusión de red neuronal	48
<i>Tabla 23.</i> Resultados del modelo de red neuronal.....	48
<i>Tabla 24.</i> Variables relevantes.....	52
<i>Tabla 25.</i> Resultados PCA	56
<i>Tabla 26.</i> Tabla de frecuencias de la variable rendimiento de pago	65
<i>Tabla 27.</i> Tabla de frecuencias de la variable capital	65
<i>Tabla 28.</i> Tabla de frecuencias de la variable rentabilidad.....	67
<i>Tabla 29.</i> Tabla de frecuencias de la variable tiempo de actividad	67
<i>Tabla 30.</i> Tabla de frecuencias de la variable satisfacción empresarial	69
<i>Tabla 31.</i> Tabla de frecuencias de la variable tiempo de actividad	69
<i>Tabla 32.</i> Tabla de frecuencias insolvencia.....	70
<i>Tabla 33.</i> Tabla de frecuencias de la variable tiempo de actividad	71

<i>Tabla 34.</i> Matriz de confusión del modelo optimizado	72
<i>Tabla 35.</i> Resultados del modelo optimizado	73
<i>Tabla 36.</i> Ajuste hiperparámetros	75
<i>Tabla 37.</i> Resultados de la optimización bayesiana	76
<i>Tabla 38.</i> Comparación de optimización	76
<i>Tabla 39.</i> Prueba de normalidad	78
<i>Tabla 40.</i> Prueba con rangos Wilcoxon	78
<i>Tabla 41.</i> Prueba no paramétrica Wilcoxon	79
<i>Tabla 42.</i> Índice de precisión con los algoritmos evaluados	79
<i>Tabla 43.</i> Índice de exactitud con los algoritmos evaluados	80
<i>Tabla 44.</i> Índice de sensibilidad con los algoritmos evaluados	80
<i>Tabla 45.</i> Índice de f1-score con los algoritmos evaluados	80
<i>Tabla 46.</i> Valoración de concordancia	82
<i>Tabla 47.</i> Medidas: Kappa de Cohen – XGBoost.....	82
<i>Tabla 48.</i> Medidas: Kappa de Cohen – SVM.....	83
<i>Tabla 49.</i> Medidas: Kappa de Cohen – LSTM	83
<i>Tabla 50.</i> Medidas: Kappa de Cohen – Bosques aleatorios.....	83
<i>Tabla 51.</i> Medidas: Kappa de Cohen – Regresión logística	84
<i>Tabla 52.</i> Medidas: Kappa de Cohen – Red neuronal.....	84
<i>Tabla 53.</i> Prueba de normalidad	85
<i>Tabla 54.</i> Correlación entre rendimiento y capital	86
<i>Tabla 55.</i> Prueba de normalidad	86
<i>Tabla 56.</i> Correlación entre rentabilidad con tiempo de actividad	87
<i>Tabla 57.</i> Prueba de normalidad	87
<i>Tabla 58.</i> Correlación entre satisfacción y tiempo de actividad.....	88
<i>Tabla 60.</i> Correlación entre insolvencia y tiempo de actividad.....	89
<i>Tabla 61.</i> Evaluación con métricas de performance.....	90
<i>Tabla 62.</i> Medidas: Kappa de Cohen – Red neuronal (MO2A)	92
<i>Tabla 63.</i> Comprobación de las hipótesis específicas	93

Lista de figuras

<i>Figura 1.</i> Arquitectura del modelo optimizado (MO2A).	27
<i>Figura 2.</i> Las características propias de los datos	28
<i>Figura 3.</i> Etapas del preprocesamiento.	29
<i>Figura 4.</i> Curva ROC XGBoost	41
<i>Figura 5.</i> Curva ROC SVM	42
<i>Figura 6.</i> Curva ROC LSTM	44
<i>Figura 7.</i> Curva ROC Regresión logística	46
<i>Figura 8.</i> Curva ROC Bosques aleatorios	47
<i>Figura 9.</i> Curva ROC Red neuronal	49
<i>Figura 10.</i> Curva ROC Algoritmos	50
<i>Figura 11.</i> Las correlaciones entre las variables	51
<i>Figura 12.</i> Factores con mayor incidencia	53
<i>Figura 13.</i> Impacto de factores	54
<i>Figura 14.</i> Características influyentes	54
<i>Figura 15.</i> Despliegue de características influyentes	55
<i>Figura 16.</i> Tasa de acierto del modelo durante el entrenamiento	58
<i>Figura 17.</i> Resultados factor rendimiento de pago	59
<i>Figura 18.</i> Resultados factor satisfacción empresarial	60
<i>Figura 19.</i> Resultados factor capital	61
<i>Figura 20.</i> Resultados factor insolvencia	62
<i>Figura 21.</i> Resultados factor tiempo de actividad	63
<i>Figura 22.</i> Resultados factor rentabilidad	64
<i>Figura 23.</i> Resultados rendimiento de pago con capital	66
<i>Figura 24.</i> Resultados rentabilidad con tiempo de actividad (meses)	68
<i>Figura 25.</i> Resultados satisfacción empresarial con tiempo de actividad	70
<i>Figura 26.</i> Resultados insolvencia y tiempo de actividad (meses)	71
<i>Figura 27.</i> Resultados prueba curva ROC	74
<i>Figura 28.</i> Validación cruzada KFold	75
<i>Figura 29.</i> RFE	77
<i>Figura 30.</i> Resultados de las métricas	81
<i>Figura 31.</i> Resultados de las métricas	91
<i>Figura 32.</i> Artículo científico publicado. SCOPUS Q4	119
<i>Figura 33.</i> Artículo científico publicado. SCOPUS Q4	120

<i>Figura 34.</i> Indexación en SCOPUS. Revistas Q4.....	121
<i>Figura 35.</i> Diploma de participación en conferencia internacional (ICWTNS).....	122
<i>Figura 36.</i> Código validación cruzada (KFold)	123
<i>Figura 37.</i> Código Shap.....	124
<i>Figura 38.</i> Código Lime.....	124
<i>Figura 39.</i> Código imputación KNN	125
<i>Figura 40.</i> Código outliers	126
<i>Figura 41.</i> Código optimización bayesiana.....	127
<i>Figura 42.</i> Código selección de características (RFE).....	128
<i>Figura 43.</i> Código reducción de dimensionalidad (PCA)	129

Dedicatoria

A mi familia, quienes me
apoyan en forma constante y
han creído en mí.

A mis Estudiantes, fuente
inagotable de inspiración y
motivación.

Agradecimiento

Al Dr. Ciro Rodriguez Rodriguez por todo el apoyo brindado para el desarrollo de la presente tesis. Asimismo, por compartir todo su conocimiento y experiencia.

A la Universidad Nacional Mayor de San Marcos – UNMSM y sus maestros por todo el conocimiento impartido.

A los directivos y funcionarios de la Cámara de Comercio de Lima que brindaron el apoyo necesario para la elaboración de la presente tesis.

Resumen

El objetivo del trabajo de investigación fue desarrollar un modelo de aprendizaje automático para optimizar la predicción de la quiebra de empresas. El tipo de investigación fue aplicada con enfoque cuantitativo y diseño experimental. Asimismo, la muestra estuvo conformada por 379 empresas ubicadas en Lima Metropolitana, donde se realizó la aplicación de un instrumento (encuesta) y se clasificó en dos dimensiones: Entorno Empresarial y Situación Financiera. En ese marco, las actividades realizadas permitieron generar la matriz del conjunto de datos de 7,959 celdas. Posterior a ello, se ejecutaron 4 etapas, la primera etapa fue realizar el preprocesamiento: Limpieza de datos, normalización y categorización. La segunda etapa fue evaluar algoritmos y seleccionar el de mejor performance. La tercera etapa consistió en implementar el modelo mediante la identificación de factores relevantes y generar correlaciones. La cuarta etapa fue evaluar el modelo mediante métricas de performance. En ese sentido, los resultados mostraron que el modelo optimizado de aprendizaje supervisado (MO2A) obtuvo precisión 97%, exactitud 97%, sensibilidad 99% y f1-score 98%. Asimismo, para comprobar las hipótesis se realizaron las medidas estadísticas de Wilcoxon, coeficientes de correlación de Rho Spearman y Kappa de Cohen. Finalmente, se precisa que el desarrollo de un modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas propicia mejores panoramas para la toma de decisiones y evitar escenarios de quiebra empresarial.

Palabras clave: Modelo, aprendizaje automático, quiebra, empresas.

Abstract

The objective of the research work was to develop a machine learning model to optimize the prediction of business bankruptcy. The type of research was applied with a quantitative approach and experimental design. Likewise, the sample consisted of 379 companies located in Metropolitan Lima, where an instrument (survey) was applied and classified into two dimensions: Business Environment and Financial Situation. In this framework, the activities carried out allowed the generation of the data set matrix of 7,959 cells. After that, 4 stages were executed, the first stage was to perform the preprocessing: Data cleaning, normalization and categorization. The second stage was to evaluate algorithms and select the one with the best performance. The third stage consisted of implementing the model by identifying relevant factors and generating correlations. The fourth stage was to evaluate the model using performance metrics. In this sense, the results showed that the optimized supervised learning model (MO2A) obtained precision of 97%, accuracy of 97%, sensitivity of 99% and f1-score of 98%. Likewise, to test the hypotheses, the statistical measures of Wilcoxon, Rho-Spearman correlation coefficient and Cohen's Kappa coefficient were performed. Finally, it is noted that the development of a machine learning model to improve the prediction of business bankruptcy provides better scenarios for decision-making and avoids business bankruptcy scenarios.

Keywords: Model, machine learning, bankruptcy, companies.

I. INTRODUCCIÓN

1.1. Situación Problemática

Las empresas se encuentran utilizando herramientas y estrategias empresariales con tecnología emergente de una manera más recurrente en un entorno empresarial muy agresivo. No obstante, en esas condiciones la generación de valor de los productos y servicios es la mayor aspiración del negocio. Asimismo, una de las principales estrategias aplicadas es el pronóstico de situaciones y eventos que podrían comprometer las actividades propias de subsistencia o de crecimiento. Por ello, es importante el desarrollo e implementación de estrategias tecnológicas en actividades cruciales y que determinan la razón del negocio. En ese contexto, la inteligencia artificial es una ciencia que ha venido realizando propuestas e iniciativas a empresas que se encuentran principalmente en subsistencia o que requieren una mayor eficiencia y optimización de sus procesos. Asimismo, se considera que los resultados que requiere una empresa se encuentran alineados a las metas institucionales y planes estratégicos.

El autor Calabrese (2023) consideró que los modelos de aprendizaje automático tienen diversos algoritmos que permiten la predicción y el pronóstico de actividades con trascendencia para la empresa. Sin embargo, es indispensable el tratamiento con criterios relevantes, variables idóneas y datos consistentes. Asimismo, los autores Lee et al. (2019) precisaron que estudios sobre pronósticos y predicciones con empresas utilizaron modelos de clasificación crediticia, rendimiento de pago y quiebra empresarial, los cuales proporcionan técnicas y factores determinantes mediante patrones específicos con el fin de generar predicciones según los criterios de selección.

Los autores Zhang et al. (2022) consideraron que las empresas enfrentan desafíos determinantes debido a las grandes dificultades financieras, la información financiera incompleta y las restricciones de capital. Por lo tanto, empresas (micro y pequeñas) suelen quedar excluidas en los canales de financiación tradicionales. En ese sentido, los desafíos financieros influyen negativamente en la salud empresarial y la sostenibilidad del desarrollo de estas empresas. Asimismo, los autores Muñoz-Cancino et al. (2023) determinaron que las empresas recién iniciadas en operaciones tienen inconvenientes en las evaluaciones crediticias lo que origina altas tasas de interés y desvirtúan la naturaleza de la financiación, la mayoría se encuentra en etapa de subsistencia. En ese sentido, las empresas no logran obtener el acceso a los servicios financieros y una evaluación crediticia adecuada. Por este motivo, a nivel macroeconómico, la mayoría de las empresas emergentes no constituyen el motor importante para el crecimiento económico local, afrontan crisis y una posible quiebra. Adicionalmente a ello, los autores Muñoz-Cancino et al. (2023) consideraron que las entidades financieras juegan un importante papel social a la hora de facilitar el acceso al crédito y afrontar los riesgos que conlleva prestar dinero. Sin embargo, las empresas experimentan una serie de dificultades para el financiamiento. Para gestionar este riesgo de crédito, las instituciones financieras han aplicado modelos de calificación crediticia para evaluar la solvencia de sus prestatarios que no son los adecuados. Para la calificación crediticia las instituciones financieras a menudo usan información personal, datos bancarios e historial de pagos, en un modelo de calificación crediticia, para estimar la solvencia y la probabilidad de incumplimiento. A pesar de ser el mecanismo estándar en la industria para las decisiones de otorgamiento de crédito y pronóstico de préstamo, esta herramienta aún no asegura el acceso adecuado al crédito y al sistema financiero.

Los autores Sulastri y Janssen (2022) consideraron que las empresas tienen dificultades en el acceso a financiamiento y la apertura a otros mercados, motivando que se generen actividades de subsistencia, los cuales con el tiempo se pueden considerar como una potencial quiebra. Del mismo modo, el autor Ben-Sassi (2022), precisó que el riesgo crediticio de las empresas en la financiación préstamos es definido como la probabilidad de que incumpla los préstamos

derivados y por lo tanto obtenga una calificación negativa. En ese sentido, se presentan datos de estados financieros que generan un pronóstico sobre riesgo para acceder al crédito de las empresas. Sin embargo, los datos de comportamiento y rendimiento permiten calificar para préstamos mediante convenios, pero una vez obtenido el préstamo es muy probable que no logren cumplir con la obligación de pago.

Según Muthukumaran y Hariharanath (2022), las empresas experimentan diversos escenarios de competitividad en mercados agresivos y complejos. Asimismo, los años de operatividad es un factor que ayuda a conocer la orientación del mercado interno y externo, pero no es una garantía de subsistencia. Asimismo, los autores Kanzari et al. (2022), precisaron que la medición del desempeño es necesario para los procesos de control y es imprescindible medir los aspectos financieros de las empresas antes y después de conseguir un préstamo. Asimismo, consideraron que la calificación del desempeño financiero origina la rentabilidad de una empresa y su continuidad a largo plazo. Según Chávez-Zavaleta et al. (2024), en los últimos años las MYPES del Perú han podido acceder a préstamos por medio de Instituciones; sin embargo, la MYPES tienen dificultades de pago luego de obtener el préstamo. En ese sentido, la calificación es negativa y no podrá acceder a préstamos futuros.

Los autores Gao y Balyan (2022) precisaron que la propuesta de un modelo de predicción de riesgos financieros y de quiebra se ha convertido en la necesidad del momento debido a los inconvenientes de largo y corto plazo en el mercado financiero para reducir y mitigar el riesgo que genera el incumplimiento de sus obligaciones. Por ello, los autores Zhang et al. (2022) consideraron que el riesgo sobre el crédito de MYPES se define como la probabilidad de que la MYPE incumpla los préstamos derivados de la financiación. Los modelos tradicionales utilizan simplemente los datos estáticos de las MYPES. Sin embargo, estos modelos pasan por alto los datos de comportamiento, que reflejan el comportamiento financiero dinámico, el cual lo que limita los cálculos sometidos a pronósticos del riesgo crediticio.

El autor Calabrese (2023) precisó que los modelos de calificación crediticia para las MYPES no reflejan oportunamente la capacidad de pago de empresas de empresa emergentes con menos de 12 meses de operatividad. La crisis financiera mundial ha demostrado la importancia de tener un modelo para predecir el incumplimiento de pago. En ese contexto, los autores Zhang et al. (2022) precisaron que lo necesario para la predicción son los datos del entorno interno, con énfasis en las finanzas. Asimismo, consideraron que el impago de los préstamos genera un impacto significativo en toda la cadena del mercado y motiva a que las entidades financieras no generen préstamos a las MYPES. Los cuales sin apoyo financiero no pueden invertir, quedándose sin efectivo y terminando en la bancarrota. En ese sentido, es necesario el desarrollo de un modelo que permita mejorar la predicción de la quiebra de empresas.

1.2. Formulación del Problema

1.2.1. Problema General

¿De qué manera un modelo de aprendizaje automático optimizará la predicción de la quiebra de empresas?

1.2.2. Problemas Específicos

- ¿Cómo preprocesar la data colectada de empresas para optimizar la predicción del modelo de aprendizaje automático?
- ¿Cómo evaluar el performance de los algoritmos de aprendizaje automático para optimizar la predicción del modelo de la quiebra de empresas?
- ¿Cómo correlacionar los factores relevantes de las empresas para optimizar el diseño del modelo de aprendizaje automático de predicción de la quiebra?
- ¿Cómo evaluar con métricas de performance el modelo de aprendizaje automático para la predicción de la quiebra de empresas?

1.3. Justificación teórica

La justificación teórica reviste de una importancia relevante debido a que se dio énfasis en los modelos conceptuales y la incorporación de variables sobre el contexto empresarial. Además, la parte teórica de fundamentación permitió identificar las relaciones entre variables, los cuales propiciaron mejores resultados. Adicionalmente a ello, las definiciones y técnicas incorporadas fueron necesarias e importantes para complementar el planteamiento teórico dentro de un modelo conceptual propicio para la optimización de los procesos y la generación de resultados mediante técnicas y métodos de desarrollo.

1.4. Justificación práctica

La justificación práctica se refiere al uso de un modelo seleccionado, optimizado y evaluado de aprendizaje automático para contribuir a la predicción de la quiebra de empresas, el cual fue importante en la generación de panoramas seguros para inversiones futuras y generar valor a las expectativas a futuro; de esta manera, las empresas ejecuten procesos más rápidos que permitan anticiparse a escenarios de quiebra y mitigar la incertidumbre. Asimismo, es importante el desarrollo de actividades de consistencia de los datos mediante técnicas adecuadas.

1.5. Objetivos

1.5.1. Objetivo general

Desarrollar un modelo de aprendizaje automático para optimizar la predicción de la quiebra de empresas.

1.5.2. Objetivos específicos

- Preprocesar la data colectada de empresas para optimizar la predicción del modelo de aprendizaje automático.
- Evaluar el performance de los algoritmos de aprendizaje automático para optimizar la predicción del modelo de la quiebra de empresas.
- Correlacionar los factores relevantes de las empresas para optimizar el diseño del modelo de aprendizaje automático de la predicción de quiebra.
- Evaluar con métricas de performance el modelo de aprendizaje automático para la predicción de la quiebra de empresas.

1.6. Hipótesis

1.6.1. Hipótesis general

El desarrollo de un modelo de aprendizaje automático optimiza la predicción de la quiebra de empresas.

1.6.2. Hipótesis específicas:

- El preprocesamiento de la data colectada de empresas optimiza significativamente la predicción del modelo de aprendizaje automático.
- La evaluación del performance de los algoritmos de aprendizaje automático optimiza significativamente la predicción del modelo de la quiebra de empresas.
- La correlación de los factores relevantes de las empresas optimiza significativamente el diseño del modelo de aprendizaje automático de la predicción de quiebra.
- La evaluación con métricas de performance del modelo de aprendizaje automático valida significativamente la predicción de la quiebra de empresas.

II. MARCO TEÓRICO

2.1. Marco filosófico o epistemológico de la investigación

Desde el marco filosófico se considera la búsqueda del objetivo generado mediante la comprensión de la realidad bajo la observación y el trabajo con herramientas idóneas, utiliza un enfoque realista y pragmático. Respecto al marco epistemológico se considera la validación de una hipótesis que fue derivada de un modelo determinado; donde precisa un análisis exhaustivo de las etapas y la validación. En ese sentido, los aportes filosóficos y epistemológicos generan decisiones que trascienden en la investigación como son: la selección de datos y construcción del conjunto de datos adecuado, la identificación de factores clave, la validación del modelo y la interpretación de los resultados. Los autores Wang et al. (2024) fundamentaron que los dos marcos presentan un complemento entre ellos, lo que permite la comprensión del uso de tecnología para resolver problemas complejos y generar conocimiento de naturaleza práctica y eficiente.

2.2. Marco filosófico

Dentro de un marco filosófico asociada a la fundamentación ontológica se considera que la naturaleza de la realidad es la base para la construcción de la investigación; se precisa que dentro de esta realidad se encuentran las motivaciones y la razón por la que el aprendizaje automático es útil para predecir la quiebra empresarial. Asimismo, estas motivaciones se asocian al realismo científico donde se consideró que los datos recopilados (para cada una de las variables) han reflejado la realidad objetiva del problema identificado; los cuales se generaron mediante relaciones y

patrones específicos. Los autores Wang et al. (2024) consideraron que para brindar soluciones reales y próximas a la realidad objetiva de una investigación se requiere el uso de herramientas idóneas. Del mismo modo, el autor Deichmann (2024) precisó que el empirismo lógico (positivismo) es un enfoque que se caracteriza por la construcción (desde la observación) de un conjunto de datos históricos de naturaleza medible y que son necesarios para el procesamiento con algoritmos. Asimismo, los autores Wang et al. (2024) precisaron que este enfoque se caracteriza por el tratamiento con métodos cuantitativos para la identificación de correlaciones y pronósticos. Respecto al pragmatismo se considera que este enfoque motiva la investigación desde el punto de vista de la utilidad; es decir, se presenta la posibilidad de analizar si las predicciones realizadas por el modelo optimizado brindarán mejores prácticas y aportes a los usuarios finales. El autor Deichmann (2024) consideró que el éxito a nivel práctico es la meta tangible en la justificación filosófica. Los autores Wang et al. (2024) consideraron que los estudios realizados sobre inteligencia artificial han mejorado la calidad de comprensión de los sistemas tradicionales; sin embargo, los factores sociales, económicos y ecológicos presentan mucha complejidad dentro de la interacción con un sistema. Por ello, los enfoques tradicionales experimentan una tendencia a necesitar otro tipo de paradigma y enfoque.

2.3. Marco epistemológico

Dentro de los fundamentos de la investigación se encuentra el proceso de construcción del conocimiento asociado a la investigación; en ese contexto, el marco epistemológico conceptualiza la forma como el conocimiento ha ido contribuyendo a la solución del problema identificado y la manera en que los resultados obtenidos son validados y presentan calidad. Los autores Pérez-Escobar y Sarikaya (2024) consideraron que el razonamiento inductivo identifica factores, relaciones y tendencias mediante el uso de datos adecuados; sin embargo, es necesario que para el entrenamiento los datos presenten una consistencia que permita generar nuevos patrones y una orientación adecuada

hacia un resultado óptimo; asimismo, es importante que las bases conceptuales en la construcción de variables se encuentren generadas de manera precisa y corresponda al problema identificado. Respecto al planteamiento de la hipótesis, los autores Wang et al. (2024) consideraron que los modelos desarrollados requieren de la formulación de hipótesis adecuadas para tener una mejor evaluación y una comprobación mediante técnicas estadísticas. Las hipótesis generadas requieren de un alcance efectivo y motive una comprobación sólida. El autor Deichmann (2024) consideró que el constructivismo se ve asociado en la interpretación de resultados, pero requiere de buenas prácticas en la selección, obtención de calidad y las decisiones del ser humano (que pueden ser subjetivas).

2.4. Antecedentes de investigación

Para obtener antecedentes sobre la investigación se realizó un estudio de revisión sistemática de la literatura con aspectos propios del contexto empresarial y los objetivos planteados. En ese marco, se obtuvo lo siguiente:

- Calabrese (2023) en su estudio: "Contagion effects of UK small business failures: A spatial hierarchical autoregressive model for binary data", precisó que las MYPES representan el 69.9% de todas las empresas del Reino Unido, donde el 60% se encuentra en el sector privado y contribuyen al 52% de toda la facturación. Asimismo, indica que el número de MYPES ha crecido un 64%. En ese sentido, propuso un sistema experto para que las instituciones puedan evaluar el riesgo crediticio y de quiebra, el cual fue diseñado específicamente para las MYPES. En el mencionado estudio se consideraron 4 principales métricas: precisión-85%, exactitud-90%, sensibilidad-92% y f1-score-89%. Se concluyó que las MYPES enfrentan factores externos diversos y que en muchos casos no se pueden controlar a tiempo, por la cual requieren de un seguimiento oportuno para no perder la calificación crediticia y el índice de quiebra.

- Lee et al. (2019) en su estudio: “Deep learning-based corporate performance prediction model considering technical capability”, realizaron el desarrollo de un modelo sobre una red neuronal que generaba la predicción de precios de las acciones utilizando índices económicos y financieros como predictores. Además, compararon el rendimiento de predicción de su modelo propuesto con el de un modelo de predicción basado en análisis discriminante múltiple. Asimismo, reforzó un estudio preliminar para predecir la quiebra de empresas utilizando un algoritmo de red neuronal, donde utilizaron los indicadores de desempeño financiero pasados de las empresas como predictores en el modelo. Además, mejoraron la precisión de predicción del modelo al adoptar un procedimiento de preprocesamiento, utilizando un enfoque de conjunto aproximado para obtener una tabla de información reducida. Consideraron 4 principales métricas: precisión-94.5%, exactitud-92.7%, sensibilidad-93.5% y f1-score-94%. Finalmente, el modelo permitió predecir el desempeño financiero de las empresas utilizando una red neuronal artificial con algoritmos de retropropagación.
- Zhang et al. (2022) en su estudio: “Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data”, precisaron la importancia del índice financiero para la predicción del riesgo de quiebra de MYPES con un modelo de bosque aleatorio mejorado, donde se hace uso de la razón financiera y los datos del contexto interno. Asimismo, la desventaja de estos modelos es que los datos que utilizaron solo incluyen la información estática de las MYPES donde el modelo predictivo que adoptaron solo permitió una única modalidad. Consideraron 4 principales métricas: precisión-88%, exactitud-89%, sensibilidad-90% y f1-score-88%. Adicionalmente a ello, consideraron el desarrollo de la metodología KDD.
- Bussmann et al. (2021) en su estudio: “Explainable machine learning in credit risk management”, precisaron que las empresas con ganancias altamente inconsistentes experimenten un flujo de efectivo bajo, lo que aumenta la probabilidad de no pagar a los acreedores y no cumplir con los

compromisos financieros. En el modelo propuesto consideraron la precisión de un modelo de red neuronal para generar resultados de predicción de riesgo de quiebra, crédito y rendimiento de pago. Asimismo, se realizaron las pruebas con los datos de calificación crediticia de empresas en Australia, donde demostraron que la red neuronal profunda se desempeñó con mayor precisión en la predicción del riesgo crediticio que otros modelos. La propuesta fue un modelo de aprendizaje profundo para generar resultados y mejorar la gestión con los riesgos financieros, y los experimentos con datos comerciales. Consideraron 4 principales métricas: precisión-94%, exactitud-94%, sensibilidad-93% y f1-score-94%. En este estudio se demostró la superioridad del aprendizaje profundo sobre otras técnicas de aprendizaje automático y la utilización de la metodología KDD.

- Balcaen y Ooghe (2019), en su estudio: “Alternative methodologies in studies on business failure: do they produce better results than the classical statistical methods”, propusieron un enfoque novedoso que emplea la red neuronal de propagación inversa (BPNN) para predecir el riesgo crediticio y de quiebra de las MYPES en China. Los experimentos en un conjunto de datos de 192 MYPES automotrices chinas demostraron que el enfoque propuesto supera a la máquina de vectores de soporte. Consideraron 4 principales métricas: precisión-93%, exactitud-92%, sensibilidad-92 y f1-score-93%. En este estudio se demostró la superioridad del aprendizaje profundo sobre otras técnicas de aprendizaje automático y basadas en reglas.
- Cantarero et al. (2019), en su estudio: “Efectos de la crisis económica sobre el emprendimiento en empresas de economía social en España: un análisis espacial”, propusieron un modelo de regresión logística que permitía la predicción del riesgo de quiebra y crediticio para las MYPES, utilizando los datos de ratios financieros y datos de gobierno corporativo utilizando un conjunto de datos de 934 pequeñas empresas. Se concluye que el 70% de empresas han obtenido una mayor calificación para futuros préstamos y el 60% han logrado identificar si se encuentran en riesgo de quiebra.

Consideraron 4 principales métricas: precisión-90%, exactitud-91%, sensibilidad-90% y f1-score-90%. Utilizaron la metodología CRISP-DM para el tratamiento y desarrollo de las tareas.

- Han y Gao (2019), en su estudio: “A chain multiple mediation model linking strategic, management, and technological innovations to firm competitiveness”, propusieron un modelo de bosque aleatorio mejorado mediante el uso de la relación financiera y de transacciones para la predicción del riesgo de quiebra en MYPES. Los resultados experimentales en un conjunto de datos de 54,390 MYPES británicas demostraron que el modelo propuesto puede mejorar significativamente la predicción del riesgo crediticio en comparación con otras predicciones. Consideraron 4 principales métricas: precisión-85%, exactitud-87%, sensibilidad-86% y f1-score-84%. Utilizaron la metodología KDD para generar datos estandarizados y resultados idóneos.
- Ravisankar et al. (2020), en su estudio: “Detection of financial statement fraud and feature selection using data mining techniques”, propusieron un modelo de regresión logística para predecir el riesgo de quiebra empresarial de las MYPES utilizando la razón financiera y datos de sentencias judiciales. Los resultados experimentales en un conjunto de datos de 1,091 MYPES chinas mostraron que el modelo propuesto supera los ratios crediticios de insolvencia. Consideraron 4 principales métricas: precisión-88%, exactitud-86%, sensibilidad-87.5% y f1-score-86.4%.
- Yu et al (2019), en su estudio: “Ensemble KNNs for bankruptcy prediction”, propusieron un modelo de fusión adaptativo de dos etapas para datos heterogéneos de múltiples fuentes (datos demográficos, estadísticos de microblogs y de transacciones de pago) para predecir el riesgo de quiebra. Consideraron 4 principales métricas: precisión-82%, exactitud-83%, sensibilidad-81% y f1-score-81%. Utilizaron un método híbrido entre KDD y las técnicas de exploración de datos, propiciando un enfoque para mejorar la consistencia e implementación del modelo.

- Kuster et al. (2023), en su estudio: “Early Insolvency Prediction as a Key for Sustainable Business Growth”, propusieron la aplicación de un sistema de diagnóstico temprano financiero para evaluar la insolvencia de las empresas mediante una red neuronal artificial. Además, consideraron las métricas de precisión, exactitud, sensibilidad y f1-score; en ese contexto, se obtuvieron los siguientes resultados: 91%, 90%, 89% y 91% respectivamente, los cuales lograron identificar resultados previos que mejoraron la predicción de la quiebra de empresas. El método utilizado fue de exploración de datos y consideraron 100 empresas Serbias entre los años 2019 y 2021.

2.5. Bases Teóricas

2.5.1. Quiebra empresarial

Ravisankar et al. (2020) consideraron que la quiebra empresarial es una etapa trascendente para una institución y que genera diversas actividades definitorias en un panorama de infortunio y desgaste. Asimismo, precisaron que la persona o cuadro directivo no tiene herramientas para cumplir con las obligaciones financieras por no contar con recursos económicos; de esta manera, solo se motivan con actividades de subsistencia hasta el cierre definitivo. En ese marco, se precisó que la quiebra es un recurso jurídico aplicado cuando el fracaso empresarial es inminente por tener pérdidas recurrentes y las inversiones no generaron ganancias. Los autores Zhang et al. (2022) precisaron que la quiebra empresarial tiene dos escenarios: el primero se presenta con la liquidación del negocio para evitar pérdidas más drásticas y el segundo cuando la empresa tiene mayores costos operacionales que retribuciones. Asimismo, los autores consideraron que la quiebra se da luego de diversas actividades de subsistencia y es el último recurso operacional. Por ello, los autores Han y Gao (2019) precisaron que la quiebra empresarial se encuentra vinculada con la insolvencia; es decir, una empresa con

insolvencia tiene una mayor probabilidad de cerrar según el grado obtenido. En ese sentido, los autores consideraron que identificar el factor de insolvencia es importante porque brinda un panorama de la situación actual. Asimismo, la quiebra empresarial tiene diversas formas y técnicas para identificarse; sin embargo, consideraron que estas técnicas deben aplicarse según el contexto interno. Los autores Horváthová et al. (2024) consideraron que la quiebra empresarial es un evento crítico que afecta con mayor frecuencia a las empresas recién iniciadas y que ocurre cuando no pueden cumplir con las diversas obligaciones financieras. Los cuales pueden ser debido a la falta de liquidez, ingresos mínimos y deudas impagas. Asimismo, precisaron que llegar a una quiebra empresarial precisa de diversas etapas, los cuales requieren ser identificadas y tratadas de manera exhaustiva: inicio, actividades, madurez, declive y quiebra. Los autores propusieron implementar procesos de mejorar y actividades seguimiento y control para identificar potenciales problemas o sucesos en cada etapa. Del mismo modo, los autores Xiaolei et al. (2024) consideraron que la quiebra empresarial tiene etapas complejas que no se determinan ni evalúan eficientemente y pueden ser el resultado de una combinación de factores internos de gestión y externos como el económico. Asimismo, precisan que el endeudamiento origina la insolvencia y luego es un indicador para la quiebra. Adicionalmente a ello, la capacidad de endeudamiento puede originar un ratio negativo, donde un exceso generará el incremento significativo del riesgo de insolvencia y quiebra. Los autores Kuster et al. (2023) precisaron que la insolvencia es una característica que ocurre cuando una empresa tiene problemas al generar ingresos para cubrir las obligaciones con el incumplimiento de pago de todo nivel tanto de operaciones o deudas. Identificaron que en temporadas con baja de ingresos las deudas son obligaciones insostenibles con el tiempo. Asimismo, consideraron que la insolvencia es uno de los principales factores para la quiebra empresarial. Posterior a ello, consideraron que los aspectos insostenibles tienen dos escenarios: el primero es generar ingresos para la subsistencia, donde se tiene un periodo incierto de operaciones y no se encuentra sujeto a crédito, y el segundo es generar la bancarrota o liquidación de la empresa con la respectiva declaración de quiebra. Ambos escenarios se consideran en empresas que tienen periodos cortos de inicio entre 0 a 1 año de operaciones. Sin embargo, precisaron que un gran porcentaje de empresas entre 1 a 3 años de

operaciones tienen deudas que no pueden ser cumplidas y por ello generan insolvencia.

2.5.2. Rendimiento de pago

Los autores Lee et al. (2019) consideraron que el rendimiento de pago se encuentra definido como la condición que tiene una empresa luego de efectuar sus obligaciones de pago, el cual desde la perspectiva de la empresa es el estado o atributo que se origina con la información financiera y el desempeño de la empresa. Sin embargo, los juicios cualitativos son muy subjetivos y limitados en el sentido de que las conclusiones tienen un costo significativo en términos de tiempo y dinero. Con los desarrollos tecnológicos recientes, los investigadores han comenzado a utilizar técnicas de aprendizaje automático para predecir el desempeño empresarial. Asimismo, los autores indicaron que las empresas no son tan rigurosas cuando tienen ganancias significativas y los gastos se incrementan mes a mes. Sin embargo, en momentos de crisis los retrasos en los pagos generan panoramas molestos tanto para los proveedores como para la misma empresa y no motiva el otorgamiento de nuevos créditos. Por ello, el rendimiento de pago cobra especial interés para obtener indicadores precisos a la hora de gestionar créditos y permitir la relación con otras variables.

2.5.3. Riesgo de crédito

Los autores Yao et al. (2022) precisaron que el riesgo de crédito empresarial se refiere a la generación de pérdidas causadas por la imposibilidad de una empresa de pagar sus cuentas y cumplir con sus obligaciones contractuales. Asimismo, la selección de características y el análisis de mejoras con factores clave no solo puede mejorar la predicción del riesgo crediticio empresarial sino también reducir los costos comerciales. En ese sentido, para lograr la predicción del riesgo de crédito empresarial es necesario que la estabilidad de la selección de características está relacionada con el rendimiento del modelo de predicción y la confiabilidad del soporte de decisiones. La investigación sobre la selección de

características se ha centrado principalmente en los métodos de selección de variables relevantes y según el contexto se pueden generar asociaciones y la determinación de mejores métricas. Del mismo modo, los autores Sulastrri y Janssen (2022) definieron a la calificación crediticia como la evaluación del riesgo asociado con el préstamo a una empresa o individuo. Por tanto, la calificación crediticia proporciona información sobre la probabilidad de incumplimientos en el futuro y para conocer comportamientos inesperados. Asimismo, la incapacidad de predecir el incumplimiento podría generar varios costos sociales para las entidades préstamos y la continuidad de las plataformas de pago.

2.5.4. Inteligencia Artificial

Las tecnologías emergentes son muy valoradas para las empresas en la actualidad y el uso de la inteligencia artificial tiene una importancia estratégica significativa para optimizar procesos, analizar comportamientos y generar pronósticos. En ese marco los autores Xiaolei et al. (2024) consideraron que la aplicación de herramientas de inteligencia artificial es ideal para identificar tendencias sobre comportamiento de las empresas y de esta manera generar predicciones para la obtención de mejores resultados. Asimismo, los autores Chávez-Zavaleta et al. (2024), precisaron que la inteligencia artificial tiene atributos y ventajas como: Mejor control para la gestión de procesos y requerimientos de tecnología, mejora en la automatización de procesos, optimizar y generar estrategias emergentes, generar predicciones y pronósticos, prevenir sucesos y actividades, y mejorar métricas de eficiencia, eficacia y efectividad. Los autores Horváthová et al. (2024) precisaron que la inteligencia artificial aplicada en las empresas permite generar decisiones idóneas mediante la identificación de tendencias, predicciones y pronósticos sobre un tema particular y crítico. Asimismo, consideraron que el uso de tecnologías emergentes es ideal para las empresas mediante el uso del análisis predictivo, automatización de procesos, gestión y optimización de modelos. Adicionalmente a ello, los autores Yahia et al. (2021) precisaron que la inteligencia artificial logra mejorar actividades con características recurrentes y complejas mediante el análisis predictivo sobre datos históricos y variables relevantes que permitan generar tendencias y comportamientos de demanda.

2.5.5. Aprendizaje automático

El aprendizaje automático es una categoría de la inteligencia artificial que considera el propósito de generar un aprendizaje más eficiente y efectivo mediante instrucciones precisas y complejas. Los autores Chávez-Zavaleta et al. (2024), consideraron que el aprendizaje tradicional tiene muchas variables, pero todas se enfocan en procesos sin un seguimiento adecuado o actividades que permitan crecer y mejorar. Por ello, consideraron que la identificación de patrones con atributos complejos y variables múltiples pero correlacionadas permiten generar mejores y optimizaciones. El proceso de aprendizaje considera instrucciones y criterios que se incorporan a un algoritmo que genera resultados con la capacidad entrenar y aprender para predecir comportamientos eficientes. Asimismo, los autores Xiaolei et al. (2024) mencionaron que la identificación de modelos de aprendizaje automático debe considerar etapas idóneas según el entorno del problema para que las variables incorporadas se adecúen a un entrenamiento eficiente. Además, los autores consideraron que los datos obtenidos son insumos para algoritmos complejos que se van tratando según el nivel de correlación que se presenten entre variables.

El aprendizaje automático es preciso para generar pronósticos empresariales y forma parte de estrategias para la innovación empresarial. En ese sentido, presenta resultados idóneos según los criterios, reglas y condiciones adecuadas. Asimismo, los autores los autores Raza et al. (2022) consideraron que los modelos gestionados con aprendizaje automático requieren de un conjunto de datos fiables y con consistencia efectiva para propiciar un adecuado entrenamiento; todo ello, permitirá la optimización de procesos y la mejora continua.

2.5.6. Clasificación de los modelos de predicción

Ben-Sassi (2022) precisó que los modelos de predicción son técnicas mediante procedimientos que tienen etapas diferenciadas para el tratamiento de los datos mediante los algoritmos seleccionados. Asimismo, mencionó que es importante identificar el modelo apropiado según la problemática. Todo ello garantizará que los

resultados y propuestas de solución se acerquen al 100%. Respecto a los algoritmos utilizados en un modelo de aprendizaje automático los autores Muthukumaran y Hariharanath (2022) consideraron que las redes neuronales son idóneas en planteamientos donde los datos requieren un constante aprendizaje. Asimismo, utilizaron técnicas complejas mediante el procesamiento de datos según patrones establecidos y que simulan las conexiones en el cerebro del ser humano. Los autores Muñoz-Cancino et al. (2023) consideraron que una red neuronal es utilizada para realizar actividades de predicción, clasificación y toma de decisiones. Asimismo, precisaron que se encuentran formadas por nodos interconectados que mejoran con un proceso de entrenamiento. Del mismo modo los autores Zhang et al. (2022) precisaron que los algoritmos de regresión logística utilizan técnicas estadísticas mediante instrucciones complejas para elaboración de pronósticos. El autor Calabrese (2023) precisó que el algoritmo de regresión logística requiere de actividades para su implementación, los cuales se determinan con la inicialización mediante el uso de parámetros, el cálculo de predicciones y probabilidades, las funciones de costo y las iteraciones.

Respecto a los algoritmos de bosques aleatorios, los autores Pratt et al. (2021) precisaron que son un conjunto de árboles de decisión que se agrupan mediante parámetros asociado para generar puntuaciones sobre la clasificación de todos los posibles resultados. Los autores Yao y Wang (2022) precisaron que los bosques aleatorios son árboles de decisión que fueron objeto de entrenamiento con variaciones y aleatoriedad en los datos. En ese sentido, el proceso de predicción se realiza por intermedio de la combinación de las predicciones realizadas por cada predicción que realiza el árbol entrenado.

III. METODOLOGÍA

3.1. Tipo y diseño de investigación

La investigación fue desarrollada mediante la metodología de tipo aplicada, relacionada a un enfoque cuantitativo y a un diseño experimental. Respecto al tipo de investigación aplicada los autores Reyes y Sánchez (2015), precisaron que el estudio de tipo aplicada formula teorías, conceptos y definiciones que influyen para la generación de nuevo conocimiento. Del mismo modo, los autores Hernández-Sampieri et al. (2018) consideraron que el conocimiento generado es indispensable e importante para resolver problemas de la sociedad en su conjunto y establecer escenarios propicios para la solución. Asimismo, el enfoque cuantitativo utiliza un conjunto de datos que permiten ser representados mediante modelos matemáticos y procesados con la aplicación de herramientas estadísticas para la descripción, explicación y predicción de situaciones o fenómenos identificados. En ese marco el autor Ñaupas (2018) consideró que el enfoque cuantitativo utiliza instrumentos de recolección de datos como encuestas y fichas de registro para el procesamiento y análisis de datos mediante el uso de herramientas estadísticas; los cuales permitirán la generación de resultados para la comprobación de hipótesis.

Respecto al diseño se determinó el diseño experimental de tipo pre-experimental con medición única, el cual es necesario para predecir la quiebra de empresas mediante la incorporación de variables en una base de datos para la aplicación de Aprendizaje Automático. Asimismo, según el autor Ramos-Galarza (2021) indicó que los estudios pre-experimentales reciben datos de un grupo poblacional identificado previamente, los cuales son necesarios para la aplicación de los instrumentos que permiten identificar la orientación del grupo. Del mismo modo, los autores Hernández-Sampieri et al. (2018), consideraron que la investigación pre-

experimental consiste en proporcionar un procedimiento o estímulo a un grupo poblacional específico (muestra); posterior a ello, identificar los resultados obtenidos. En la Tabla 1 se muestra el diagrama para el diseño de la investigación, donde se precisa el grupo de estudio que son las empresas, el tratamiento o estímulo y las métricas.

Tabla 1. Diagrama del diseño de investigación

G	X	O₁
G: Grupo de estudio: Empresas X: Tratamiento o estímulo aplicado: Modelo de ML O ₁ : Métricas		

Fuente. Hernández-Sampieri et al. (2018)

3.2. Unidad de análisis

El sujeto de estudio fue la empresa de la ciudad de Lima Metropolitana según criterios de inclusión y exclusión.

3.3. Población y muestra

Los autores Hernández-Sampieri et al. (2018) consideraron que la población es un grupo con características determinadas y que son descritas con criterios principales para la obtención de una muestra. Se deben incluir intervalos de tiempo, ámbitos de acción, accesos y contenido. La población de estudio se encuentra determinada por las empresas de la ciudad de Lima Metropolitana. El cual fue precisado según el documento de demografía empresarial del Perú (I trimestre de 2024) realizado por el Instituto Nacional de Estadística e Informática (INEI), y tiene como fuente el Directorio Central de Empresas y Establecimientos (DCEE). En ese sentido, se tiene 25,880 empresas dadas de alta al I trimestre del 2024 en Lima Metropolitana.

3.4. Selección de la muestra

Los autores Hernández-Sampieri et al. (2018) definieron que una muestra es el subgrupo de una población objeto de estudio que tiene representatividad sobre los resultados obtenidos hacia el grupo poblacional. La muestra de estudio fue de 379 empresas de la ciudad de Lima, Perú (microempresas, pequeñas empresas y medianas empresas). El tamaño de muestra ha sido calculado con un margen de error de 5% y el nivel de confianza de 95% utilizando la fórmula de cálculo muestral de universo finito. Asimismo, mediante un proceso de exploración se identificó 118 registros que corresponden a microempresas, 166 a pequeñas empresas y 95 a medianas empresas. La fórmula de muestra finita es:

$$M = \frac{d^2 Pab}{e^2(P - 1) + d^2 ab}$$

M = muestra

d = desviación de curva

a = probabilidad de éxito (0.5)

b = 1 – p = 0.5

P = Población = 25,880

e = 0.05 máximo error permitido

Respecto al muestreo se determinó como probabilístico porque todos los sujetos (objeto de estudio) tienen la misma posibilidad de ser elegidos. Asimismo, los autores Hernández-Sampieri et al. (2018) mencionaron que es necesario considerar que el grupo poblacional tiene criterios específicos según el contexto empresarial, los cuales son necesarios para obtener resultados y comprobar hipótesis.

3.5. Recolección de datos

Al respecto Hernández-Sampieri et al. (2018) precisaron que los cuestionarios son instrumentos de aplicación directa para conocer necesidades, percepciones,

comentarios o cuestionamientos sobre el desempeño de un determinado tema. Asimismo, los autores definen que el cuestionario es un conjunto de preguntas relacionadas con una o más variables para el tratamiento preciso de la observación realizada (p. 217). En ese marco, la encuesta es imprescindible para recopilar datos primigenios que ayudarán a constituir un dataset propicio para el tratamiento y posterior análisis que permitirán generar resultados para mejorar la predicción de la quiebra empresarial.

El instrumento fue elaborado mediante Microsoft Word y la aplicación fue realizada a todos los gerentes de las empresas. Asimismo, estuvo compuesto por 20 preguntas clasificados según la dimensión que corresponde: Entorno Empresarial (10) y Situación Financiera (10), dichas preguntas fueron elaboradas según el análisis del contexto empresarial. Asimismo, el instrumento aplicado permitió utilizar la técnica de observación para identificar las fichas de registro que son los necesarios para generar resultados de las métricas para cada algoritmo seleccionado. Los autores Hernández-Sampieri et al. (2018) definieron al método de observación como una herramienta útil para la recolección y documentación de datos mediante fichas de registro que luego son necesarios para el procesamiento de los datos.

3.6. Confiabilidad de datos

Para validar la aplicabilidad del instrumento (encuesta) para la recolección de datos y la posterior construcción del dataset se realizaron los siguientes procedimientos:

- a. Análisis y evaluación del instrumento (encuesta)
- b. Análisis y evaluación de las fichas de los 3 modelos.

En ese sentido, luego del análisis y evaluación por parte de los 3 expertos se precisó que el procedimiento a y b son favorables y aplicables para el estudio que se propuso. Respecto a la prueba de confiabilidad del instrumento (encuesta) se

determinó mediante el coeficiente de Alpha de Cronbach donde se obtuvo el valor de 0.817 que representa como “buena” y “muy alta”, en ese marco los autores Hernández-Sampieri et al. (2018) consideraron que la fiabilidad del instrumento es aquel que brinda solidez a un instrumento para ser aplicado en una población mediante una muestra. Asimismo, la investigación consideró la evaluación de 3 expertos, orientadas a la investigación y a la carrera de ingeniería de sistemas, quienes brindaron validez a los instrumentos propuestos.

Los autores Hernández-Sampieri et al. (2018) determinaron que la validez de instrumento se define como el juicio de valor (Evaluación de pertinencia e idoneidad) sobre el instrumento que medirá la variable. En ese sentido, en la Tabla 2 se precisaron los expertos especializados en los temas técnicos y de investigación, los cuales respondieron a los criterios de idoneidad y experiencia.

Tabla 2. Validez del instrumento

N°	Validador	Calificación
1	Dr. Javier Gamboa Cruzado	Aplicable
2	Dr. Alfredo Daza Vergaray	Aplicable
3	Dr. Willy Ugarte Rojas	Aplicable

Fuente. Elaboración propia.

3.7. Análisis e interpretación de datos

La investigación inició con la búsqueda de información de investigaciones similares al tema propuesto como tesis o artículos tanto a nivel nacional como internacional. El cual permitió conocer la situación actual de la quiebra empresarial. Asimismo, una vez obtenido los datos recopilados se procedió a identificar los antecedentes y precisar las bases teóricas que brindaron solidez a la propuesta y al estado del arte. Para la recolección de datos se remitió el instrumento (encuesta) a todos los Gerentes (identificados en la muestra) mediante un formulario online que contenía las 20 preguntas. Asimismo, con los datos recopilados se procedió a construir la matriz del conjunto de datos para iniciar la etapa de preprocesamiento y luego

evaluar el performance de los algoritmos. Posterior a ello, se identificó el mejor algoritmo de aprendizaje automático que fue base para el desarrollo del modelo de predicción propuesto en la investigación. Finalmente, para validar las hipótesis, se utilizaron métodos estadísticos según la hipótesis específica correspondiente: Prueba de Wilcoxon, coeficiente de correlación de Rho Spearman y el coeficiente de Kappa de Cohen, los cuales fueron necesarios para determinar si existen cambios significativos luego del preprocesamiento, evaluar cómo se desempeñaron los algoritmos de aprendizaje automático, determinar si existe relación entre factores relevantes y evaluar con métricas de performance el modelo propuesto.

Los autores Beck et al. (2024) precisaron que la prueba de Wilcoxon es necesaria para comprobar si se presentan cambios en dos muestras relacionadas. Del mismo modo, indicaron que la prueba de correlación de Rho Spearman determina el grado de relación que existe entre dos variables. Asimismo, consideraron que el coeficiente de Kappa de Cohen es necesario para medir el nivel de concordancia entre diferentes evaluaciones para obtener resultados confiables y con consistencia. Asimismo, los autores Montoya y Montoya (2011) consideraron que las pruebas para determinar la relación entre variables necesitan de factores comunes para garantizar una adecuada relación, del mismo modo se requiere que las variables consideren una adecuada consistencia de datos. Del mismo modo los autores Mobbs et al. (2024) consideraron que Kappa de Cohen es utilizado en la minería de datos para evaluar el rendimiento de algoritmos o modelos de aprendizaje automático y también para seleccionar parámetros o factores óptimos para un modelo. Adicionalmente a ello, consideraron que el coeficiente Kappa de Cohen tiene ventajas sobre otras métricas que se utilizan para evaluar modelos de aprendizaje automático, como la exactitud, precisión, la sensibilidad o F1-Score.

Las cuales en algunos eventos pueden tener resultados sesgados cuando las variables no tienen relaciones óptimas o cuando los datos comparados no tienen un atributo aleatorio. En ese sentido, Kappa de Cohen considera los factores indicados y a su vez proporciona medidas sólidas y significativas respecto a la concordancia. Finalmente, se realizaron los análisis descriptivo e inferencial para

obtener resultados indispensables para obtener una adecuada discusión de resultados con los antecedentes identificados previamente.

3.8. Dataset

El dataset se encuentra compuesto de 379 registros que correspondieron a empresas pequeñas y medianas de Lima Metropolitana. En Tabla 3 se identificó las variables, tipos y la definición de cada una de ellas.

Tabla 3. Variables, tipos y definición

Nro	Variable	Tipo	Definición
0	Quiebra	int 64	Identificador de quiebra
1	Meses_A	object	Meses de actividad que tiene la empresa
2	Capital	object	Capital inicial de la empresa
3	Planes	int 64	Planes normativos
4	SI	object	Sistemas de información
5	Calificación	int 64	Calificación crediticia
6	Capacidad_E	int 64	Capacidad de endeudamiento
7	Satisfacción_E	object	Satisfacción empresarial / Clima laboral
8	Acceso_Credito	int 64	Acceso a crédito
9	Acceso_Mercados	int 64	Acceso a mercados externos
10	Personal	int 64	Empleados que trabajan en la empresa
11	Letras_por_pagar	object	Letras pendientes de pago
12	Sociedades	int 64	Conformación de sociedad
13	Rendimiento_Pago	int 64	Rendimiento de pago
14	Competidores_D	int 64	competidores directos tiene la empresa
15	Imprevistos	object	Involucramiento en Siniestros
16	Siniestros	int 64	Presentación de imprevistos
17	Tasa_utilidades	object	Tasa de utilidades
18	Rentabilidad	int 64	Índice de rentabilidad
19	Eficiencia	int 64	Índice de eficiencia en ventas
20	Insolvencia	int 64	Índice de insolvencia

Fuente. Elaboración propia.

3.9. Modelo de aprendizaje automático propuesto (MO2A)

3.9.1. Método

Se utilizó el método de análisis exploratorio de datos que se enfoca a al examinación, resumen y visualización de los resultados de un conjunto de datos que permiten entender las características idóneas y principales. Los autores Nicodemo y Satorra (2022) precisaron que el análisis exploratorio de datos permite describir mediante herramientas visuales y aspectos estadísticos la orientación de los datos sobre un contexto propio mediante la relación de sus variables relevantes y con valores que pueden ser atípicos.

3.9.2. Modelo propuesto

Se desarrolló el modelo optimizado de aprendizaje automático (MO2A) teniendo como base el modelo propuesto por los autores Alsheref et al. (2022) quienes determinaron tres etapas: exploración de datos, selección del algoritmo e implementación. En ese contexto, se optimizaron las etapas exploración de datos como preprocesamiento y selección de algoritmo como evaluación de algoritmos; posterior a ello, se incorporaron 2 etapas, la primera etapa: Optimización del modelo con actividades complementarias, donde se incluyeron la identificación de factores (variables relevantes), las correlaciones entre ellas, el entrenamiento, la implementación y la generación de métricas. La segunda etapa: Evaluación del performance para el modelo optimizado. Al respecto el autor Trujillo (2019) consideró que la aplicación de una metodología en soluciones de aprendizaje automático permite brindar solidez a la propuesta y generar métodos adecuados para el tratamiento de los datos. En ese sentido, se consideraron 21 variables que fueron trabajadas en la etapa de exploración de datos. Para la etapa selección del algoritmo se consideraron 6 algoritmos, donde el algoritmo de red neuronal fue el que obtuvo resultados con mayores porcentajes en las métricas establecidas. Para la etapa de optimización de modelo se identificaron que las variables relevantes

fueron: Rendimiento_Pago, Capital, Letras_por_pagar, Acceso_Crédito, Satisfacción_E y Rentabilidad.

En la Figura 1 se presentó la arquitectura del modelo propuesto, donde se consideró una mejora al análisis exploratorio de datos para obtener factores útiles sobre la quiebra empresarial e identificar las principales causas que originan la quiebra. Asimismo, se utilizó gráficos, análisis de series de tiempo y correlaciones. Este tipo de análisis permite demostrar que los patrones de datos son útiles para generar un pronóstico sobre las afectaciones hacia el efecto quiebra de las empresas, donde la identificación de los factores es muy determinante. Se ha identificado que las variables Rendimiento_Pago y Satisfacción_E se encuentran correlacionadas con la variable Meses_A por lo que se precisó el estudio de estas correlaciones para conocer características y el impacto. Se obtuvo correlaciones para identificar variables que tienen incidencia directa con la quiebra empresarial.

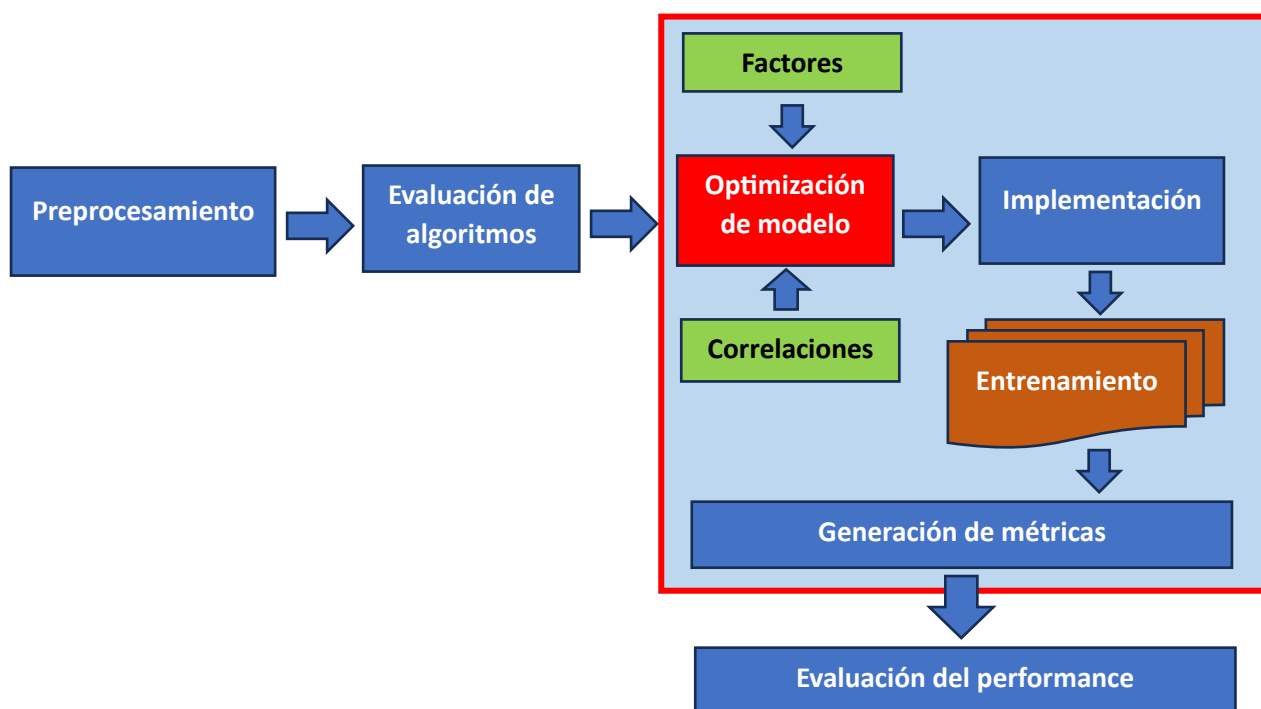


Figura 1. Arquitectura del modelo optimizado (MO2A).

Fuente. Elaboración propia.

3.9.3. Etapas del modelo propuesto

3.9.3.1. Preprocesamiento

Para la construcción del dataset se consideró 21 variables para un total de 379 registros que correspondieron a empresas pequeñas y medianas de Lima Metropolitana. Asimismo, fueron obtenidas en la etapa de recolección de los datos mediante los instrumentos requeridos como el cuestionario y las fichas de registro. En la Figura 2 se consideraron la visualización estadística de cada variable, donde se precisan 7,569 datos registrados en la matriz del conjunto de datos.

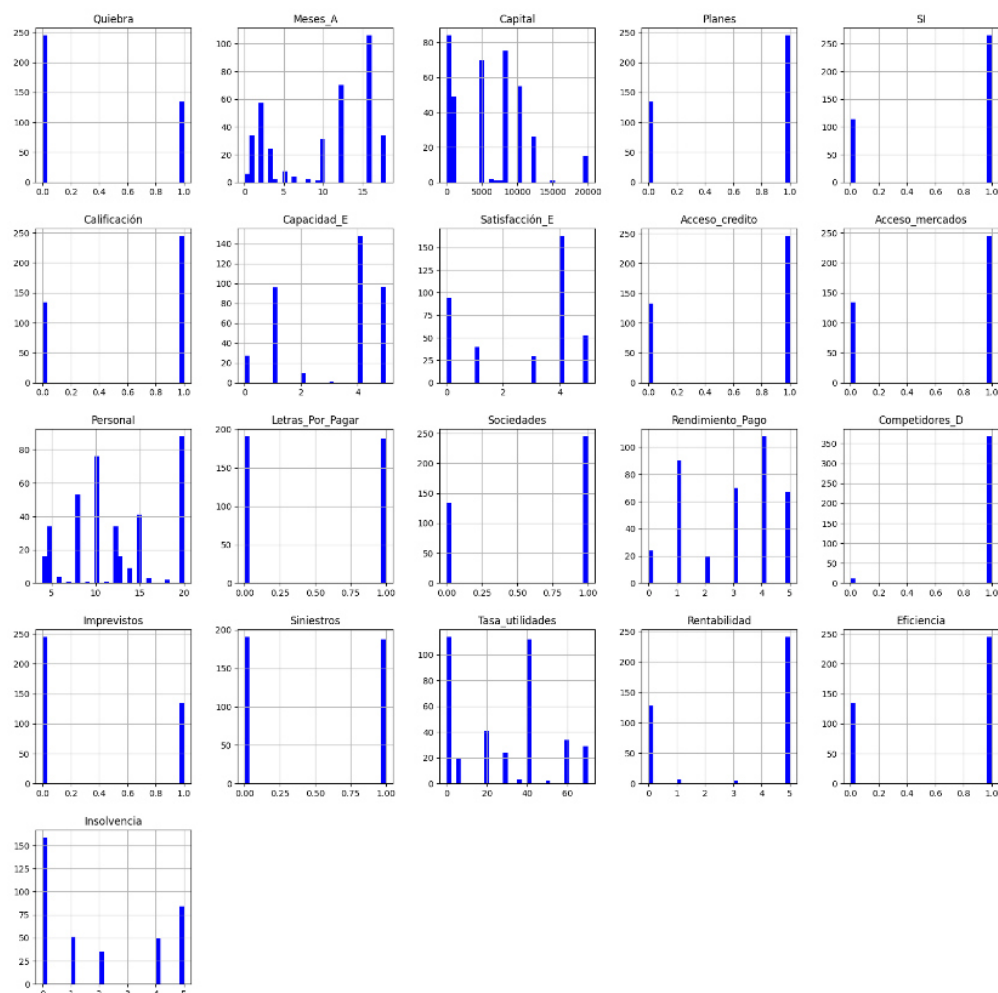


Figura 2. Las características propias de los datos

Fuente. Elaboración propia.

En ese marco, los 7,959 datos obtenidos de la matriz del conjunto de datos se refieren a las 21 variables de las 379 empresas registradas que fueron objeto de la muestra poblacional. Asimismo, se consideró la necesidad de conocer el porcentaje de consistencia para dar garantía y confianza a los resultados obtenidos; en ese sentido, fue necesario implementar la etapa de preprocesamiento donde se ejecutaron tres (3) actividades esenciales: limpieza de datos, normalización de datos y categorización de datos. Al respecto los autores Garcete et al (2019) consideraron que la técnica requerida para mejorar la consistencia de un dataset es la limpieza de los datos y para el mejoramiento se requiere de actividades normalización, el cual es necesario para transformar los datos en una estructura mucho más estable y brindarle calidad. En la Figura 3 se consideraron las etapas del preprocesamiento teniendo como base la matriz del conjunto de datos. Asimismo, se identificaron que las actividades de Limpieza, Normalización y Categorización fueron esenciales para la calidad de los resultados.

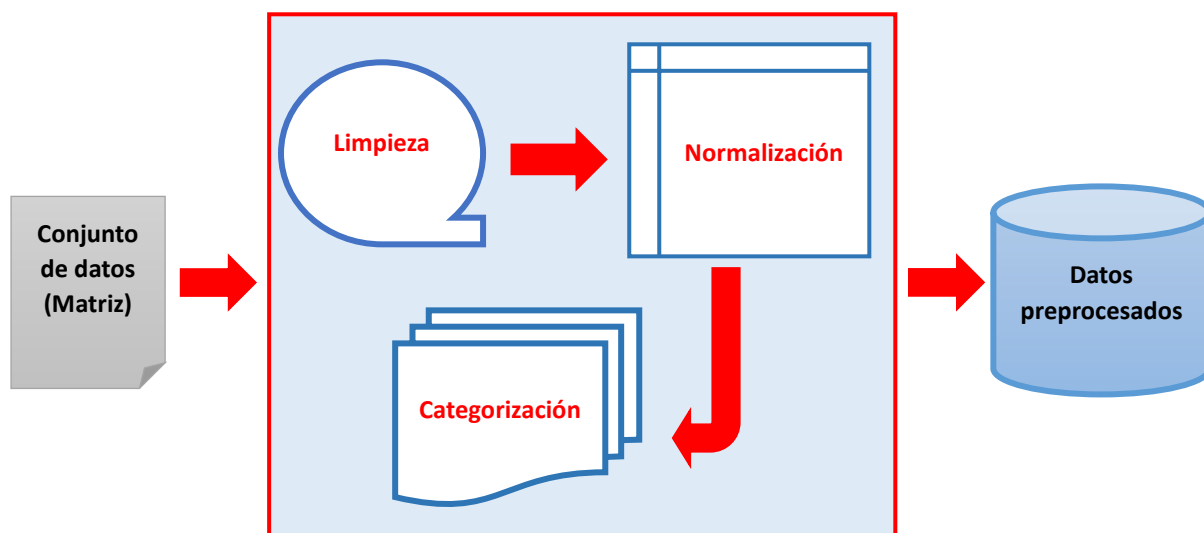


Figura 3. Etapas del preprocesamiento

Fuente. Elaboración propia.

Para la actividad limpieza de datos se realizó en base a los 7,959 datos registrados en la matriz del conjunto de datos. Asimismo, identificaron 1,235 celdas que no consignaron algún dato, 652 celdas que tienen datos con caracteres inconsistentes, 318 celdas con datos irrelevantes (registrados sin sustento ni dentro de los criterios

previamente establecidos) y 15 celdas con datos fuera del rango establecido (para el campo “Tasa” se identificó que excedían el 100%). En ese sentido, se registró el valor 0 para los casos numéricos y NP (No Precisa) para los datos de texto. Respecto al campo “Tasa” se registró el valor más alto que es 100%. No se eliminaron registros porque la matriz del conjunto de datos requiere 379 registros según el cálculo de la muestra realizada.

En la Tabla 4 se precisaron los detalles de la actividad realizada en la limpieza de datos, donde no se contempló la eliminación de algún dato inconsistente sino brindarle el tratamiento adecuado para no perder la calidad y recuperar la orientación del registro.

Tabla 4. Limpieza de datos

Celdas	Detalle	Limpieza	Tipo
1235	Sin datos	0-NP	numérico y texto
652	Caracteres inconsistentes	0-NP	numérico y texto
318	Datos irrelevantes	0-NP	numérico y texto
15	Fuera del rango (Excede el 100%)	100%	Porcentual

Fuente. Elaboración propia.

Respecto a la actividad de normalización de datos se realizó mediante la técnica de Normalización Min Max, donde se identificaron las variables: Quiebra, Planes, SI, Acceso_Mercados, Sociedades, Siniestros, Imprevistos, Acceso_Credito y Letras_por_pagar, cuyos datos registrados fueron: “SI” o “NO”; posterior a ello, se realizó el registro de la equivalencia según el rango 0 y 1, donde “NO” es 0 y “SI” es 1. Asimismo, para los datos registrados como NP (No Precisa) se generó la equivalencia y registro de 0, porque se asumió un valor mínimo dentro del rango establecido.

En la Tabla 5 se puede apreciar las variables que utilizaron una equivalencia de rango 0 y 1.

Tabla 5. Normalización Min Max

Variable	Registro	Equivalencia
Quiebra	"NO","SI"	[0-1]
Planes	"NO","SI"	[0-1]
SI	"NO","SI"	[0-1]
Acceso_Mercados	"NO","SI"	[0-1]
Sociedades	"NO","SI"	[0-1]
Siniestros	"NO","SI"	[0-1]
Imprevistos	"NO","SI"	[0-1]
Acceso_Credito	"NO","SI"	[0-1]
Letras_por_pagar	"NO","SI"	[0-1]

Fuente. Elaboración propia.

Adicionalmente a ello, se realizó el método de imputación KNN para convertir datos inconsistentes en datos válidos que son necesarios para el trabajo posterior. Los resultados obtenidos permitieron comprobar y establecer una correspondencia entre los resultados de la actividad de normalización Min Max. En ese sentido, se ingresó el número de vecinos igual a 5, el cual permitió identificar los registros similares para establecer una media. El método trabajó con la determinación de la distancia-euclídeana (número positivo obtenido en la separación de dos puntos A-B). Asimismo, los valores de texto se cambiaron a un valor binario para un mejor tratamiento de los datos. En ese sentido, en la Tabla 6 se precisaron los datos con inconsistencias identificadas y los resultados obtenidos luego de aplicar el método:

Tabla 6. Imputación KNN

Nro	Variable	Datos Totales	Datos Inconsistentes	Datos Consistentes (Imputación KNN)
0	Quiebra	379	10	0
1	Meses_A	379	78	0
2	Capital	379	67	0
3	Planes	379	98	0
4	SI	379	130	0
5	Calificación	379	96	0
6	Capacidad_E	379	185	0
7	Satisfacción_E	379	122	0
8	Acceso_Credito	379	102	0
9	Acceso_Mercados	379	139	0
10	Personal	379	189	0

11	Letras_por_pagar	379	118	0
12	Sociedades	379	155	0
13	Rendimiento_Pago	379	36	0
14	Competidores_D	379	148	0
15	Imprevistos	379	154	0
16	Siniestros	379	175	0
17	Tasa_utilidades	379	62	0
18	Rentabilidad	379	84	0
19	Eficiencia	379	40	0
20	Insolvencia	379	32	0
Totales		7,959	2,220	0

Fuente. Elaboración propia.

Se obtuvo resultados similares a lo realizado con la técnica de normalización Min Max; por lo tanto, el conjunto de datos utilizó los datos de las 7,959 celdas. Asimismo, dentro de las actividades para conseguir el 100% de consistencia de los datos fue necesario la identificación y el análisis de los valores atípicos (outliers) mediante el uso de la técnica de insolation forest, donde se obtuvo los valores mostrados en la Tabla 7.

Tabla 7. Outliers

Outliers	Valores atípicos	Recurrencia
Capital	0, 12000, 15000, 20000	<22/379
Tasa de utilidades	60, 70	<34/379

Fuente. Elaboración propia.

Se identificaron cuatro (4) tipos de valor de menor recurrencia de la variable: Capital y dos (2) tipos de valor de menor recurrencia de la variable: Tasa de utilidades. En ese sentido, se consideró que los valores si se relacionan con la naturaleza de la investigación y la variabilidad que pudo originarse a partir de ello. Asimismo, fueron parte del rango permitido en la recopilación de los datos para la conformación de la matriz del conjunto de datos.

Respecto a la actividad categorización se precisa que se realizó luego de la actividad de normalización donde se generaron dos (2) categorías: el primero fue Entorno y el segundo Situación. Estas categorías permitieron agrupar las variables según los detalles específicos de las empresas. Para la categoría Entorno se agrupó a las variables que representan las características propias de la empresa y para la categoría Situación se agruparon las variables que representan las características financieras de la empresa. Esta agrupación permitió tener un mejor panorama para los indicadores y generar correlaciones entre las variables.

3.9.3.2. Evaluación de algoritmos

Los autores Pratt et al. (2021) precisaron que el algoritmo de bosques aleatorios es idóneo para determinar la predicción de un conjunto de datos que tiene variables que puedan correlacionarse y respondan a problemas de identificación del riesgo sobre un contexto específico y detallado. Asimismo, mencionan que los algoritmos red neuronal y regresión logística pueden mejorar con entrenamiento e identificación de factores de mejora. Del mismo modo los autores Pandey et al. (2023) consideraron que para lograr métricas sólidas es importante implementar modelos con características propias del contexto analizado y que se logren alinear a las metas Institucionales. En ese contexto, los autores Yahia et al. (2021) consideraron que para obtener mejores resultados es necesario generar procesos de optimización que incluyan la incorporación de variables más idóneas y correlaciones entre ellas.

Los autores Raza et al. (2022) precisaron que los algoritmos regresión logística y bosques aleatorios usan información dinámica del mercado y variables propias del contexto empresarial. Asimismo, estos algoritmos se encuentran operativos e idóneos para estimar probabilidades de inversión y pronósticos de diversos tipos como el acceso al crédito y los pronósticos de quiebra empresarial. Del mismo modo los autores Muthukumaran y Hariharanath (2022) precisaron que el aprendizaje profundo tiene mejor interacción y se encuentra óptimo para soluciones con MYPES, debido a su versatilidad y la capacidad de generar iteraciones de entrenamiento. El principal objetivo de la técnica es determinar el estado financiero

de las MYPES, el cual contiene el diseño de la selección de características basada en el algoritmo de optimización y red neuronal, el cual se utiliza para la clasificación de datos. Las métricas son necesarias y conforman un papel importante para definir la rentabilidad de las empresas financieras y la productividad.

Para seleccionar un modelo de aprendizaje automático se requiere que los objetivos del estudio estén alineados con las actividades a desarrollar. Asimismo, es importante contar con el detalle de las variables de contexto generales y específicas. En este sentido, para la selección del modelo se realizaron pruebas de aplicación sobre el conjunto de datos con 3 modelos de aprendizaje automático que responden a los objetivos del estudio. Los cuales son: Regresión logística, bosques aleatorios y redes neuronales.

El autor Calabrese (2023) precisó que la regresión logística es útil en análisis de predicciones y reducción de costos empresariales, además para generar aumentos de eficiencia y mejora continua. Las empresas pueden descubrir patrones para mejorar la retención de los empleados, quiebra empresarial o la rentabilidad. Asimismo, consideraron que las características indispensables de este algoritmo son: Identificar la pregunta, recopilar datos históricos y realizar el entrenamiento. Los autores Yao y Wang (2022) precisaron que las técnicas realizadas por bosques aleatorios son importantes porque generan un solo resultado luego de la combinación de las predicciones realizadas por los árboles de decisión. Asimismo, la facilidad de aplicación y flexibilidad ha logrado implementarse con éxito, ya que maneja problemas de clasificación y regresión a un nivel general y específico. Asimismo, los árboles de decisión suelen tener mejor potencia en el resultado cuando las variables se encuentran correlacionadas. Los autores Muñoz-Cancino et al. (2023) consideraron que las redes neuronales son idóneas para soluciones que generen pronósticos porque generan correlaciones desde identificación de las variables, las cuales son necesarias para establecer puntos de enfoque para lograr un entrenamiento efectivo y generar el aprendizaje preciso para obtener resultados idóneos. En ese sentido, las redes neuronales consideran datos de entrenamiento para aprender y propiciar la mejora de la precisión con el tiempo. Al respecto, el modelo de optimización propuesto considera que el entrenamiento es un proceso

indispensable para mejorar los resultados de las métricas de un conjunto de datos específico que considere factores relevantes y correlaciones. Por ese motivo, se consideró realizar la evaluación de los algoritmos mediante una puntuación sobre el contexto del conjunto de datos y los requerimientos de la solución.

En la Tabla 8 se aprecia la comparación de los diversos algoritmos para el tratamiento del problema específico, donde se identificó las ventajas y desventajas:

Tabla 8. Compararon diversos algoritmos

Nro	Algoritmo	Ventajas	Desventajas	Puntuación
1	Regresión Lineal	Facilidad para el entendimiento preciso y claro en el despliegue.	No da soporte para relaciones complejas. Tendencia a sobre ajustes.	2
2	Regresión Logística	Facilidad para el entendimiento preciso y claro en el despliegue. Soporte para generar relaciones entre variables.	Las relaciones complejas no consideran todas las características. Requiere complementos para mejorar las relaciones complejas.	4
3	Árbol de decisión	Facilidad para el entendimiento preciso y claro en el despliegue. Se generan relaciones entre variables. Facilidad en la implementación.	Las relaciones complejas requieren un gran consumo de recursos.	3
4	Bosques aleatorios	Facilidad para el entendimiento preciso y claro en el despliegue. Soporte para generar relaciones entre variables. Combinación de factores y requisitos individuales y de grupo. Entrenamiento eficiente y efectivo. Performance alto.	Las relaciones complejas requieren un gran consumo de recursos.	4
5	XGBoost	Minimiza el error y mejora mediante iteraciones. Detección de valores faltantes. Extracción de características influyentes.	Requiere gran capacidad de memoria. La optimización resulta con complejidad con el uso de hiperparámetros. Requiere técnicas adicionales de mejora. Entrenamiento lento.	4
6	SVM	Facilidad para el despliegue con múltiples variables. Eficiente con el conjunto no es mayor. Ideal en clasificación binaria.	No tiene buena escalabilidad cuando el conjunto es amplio. No es ideal con valores atípicos	4
7	LSTM	Ideal para PNL.	Requieren memoria y disposición de tiempo.	4

		Ideal para datos con texto, series de temporalidad y reconocimiento de voz-video. Útiles para la traducción. Facilidad para el entendimiento preciso y claro en el despliegue.	No son ideal para datos con secuencias estáticas. No es usado para conjunto de gran tamaño.	
8	Aumento de gradiente	Soporte para generar relaciones entre variables. Performance alto. Facilidad para el entendimiento preciso y claro en el despliegue.	Cambios pequeños de características producirá cambios significativos.	2
9	Red neuronal	Se generan relaciones entre variables. Entrenamiento eficiente y efectivo. Performance alto.	Demora en el despliegue de resultados del entrenamiento (para algunos casos complejos).	5

Puntuación: 1: Inadecuado 2: Casi inadecuado 3: Ni inadecuado ni adecuado 4: Casi adecuado 5: Adecuado

Fuente. Elaboración propia.

Para seleccionar un algoritmo se requiere que los objetivos del estudio se encuentren alineados con las actividades a desarrollar. Asimismo, es muy importante tener el detalle de las variables de contexto general y específico. En ese sentido, se realizó la comparación de los nueve (9) tipos de algoritmos propuestos mediante la puntuación según las ventajas y desventajas de cada tipo.

3.9.3.3. Optimización del modelo

Las variables de estudio tienen atributos significativos para la identificación de características específicas de la quiebra. De esta manera la predicción sobre la quiebra se sustentó no solo el contexto general de la empresa sino también en el contexto interno de los procesos. Es importante que el conjunto de datos utilizado cuente con variables con un nivel de desagregación específico y que permita generar relaciones con otras variables. Para garantizar la eficacia del modelo elegido es necesario considerar actividades del proceso de ingeniería de características donde se establecieron dos criterios indispensables: 1. La identificación de factores relevantes, y 2. Correlación de variables, el cual fue precisado en la etapa de Selección del algoritmo (3.9.3. Etapas del modelo propuesto), donde se precisó que una de las ventajas del algoritmo de red neuronal es la generación de relaciones entre variables. Respecto al punto 1 se realizaron

las técnicas de Shap y Lime para la interpretación de los factores que obtuvieron mayor influencia en el modelo, los cuales fueron determinantes para la predicción. Adicionalmente a ello, se procedió a la reducción de la dimensionalidad mediante la técnica PCA, el cual fue idóneo para mejorar el rendimiento y reducir el tamaño de los datos sin perder calidad, donde se permitió establecer una comparación y confirmación de los resultados obtenidos. Respecto al punto 2 se realizó la identificación de variables que tienen un mayor porcentaje de correlación y que significaron un factor determinante para mejorar la predicción. Posterior a ello, se procedió a construir nuevos indicadores a partir de los factores relevantes identificados, los cuales se generan de la combinación de variables mediante una fórmula matemática y se constituyeron como indicadores de complemento. Respecto al entrenamiento y generación de métricas se consideró que para el logro de resultados idóneos y con características de optimización fue necesario que se considere al algoritmo que obtenga resultados mayores a los resultados de los otros algoritmos sujetos a evaluación. Asimismo, fue indispensable considerar que al proceso de entrenamiento como un factor clave para el aprendizaje del algoritmo; todo ello, garantizó resultados satisfactorios, eficientes y verídicos.

3.9.3.4. Evaluación del performance

En esta etapa se consideró la evaluación del modelo de aprendizaje automático optimizado con métricas de performance para validar la propuesta. Asimismo, esta etapa fue importante y trascendente porque los resultados obtenidos representan que el modelo para predecir la quiebra de empresas es el idóneo y eficiente según el contexto planteado. La evaluación se realizó mediante la comparación de los resultados de las métricas de cada algoritmo y la validación cruzada (KFold). Adicionalmente a ello, se realizó el proceso de ajuste de hiperparámetros para la optimización de los factores relevantes para un mejor rendimiento del modelo optimizado. Se utilizó la técnica de optimización bayesiana, el cual es más eficiente que otras técnicas y utiliza menos iteraciones. Asimismo, permitió la comparación de resultados obtenidos del modelo optimizado. Finalmente, se realizó el análisis de la importancia de los factores mediante la técnica de selección de características (RFE), el cual fue útil para determinar el impacto para la predicción.

IV. RESULTADOS Y DISCUSIÓN

4.1. Análisis, interpretación y discusión de resultados

4.1.1. Análisis de resultados del preprocesamiento

Respecto a los resultados de la estadística descriptiva se precisó que el conjunto de datos que fue colectada antes del preprocesamiento se consideró como línea base y luego de las actividades del preprocesamiento se consideró como datos preprocesados. Los resultados de la línea base se precisaron en la Tabla 9.

Tabla 9. Matriz de línea base

		F	P%	P% válido	P% acumulado
Válido	0	324	85,5	85,5	85,5
	1	55	14,5	14,5	100,0
	Total	379	100,0	100,0	

Fuente. Elaboración propia.

Los resultados del preprocesamiento se precisaron en la Tabla 10.

Tabla 10. Matriz de datos preprocesado

		F	P%	P% válido	P% acumulado
Válido	0	308	81,3	81,3	81,3
	1	71	18,7	18,7	100,0
	Total	379	100,0	100,0	

Fuente. Elaboración propia.

Se generaron resultados de antes y después de las actividades en la etapa del preprocesamiento, los cuales determinaron que si hubo una marcada diferencia. En la siguiente Tabla 11 se puede apreciar que la matriz del conjunto de datos precisó el 28% de celdas inconsistentes (72% de consistencia); motivo por el cual, se aplicaron actividades de preprocesamiento que se precisó en la etapa: 3.9.3.1. Preprocesamiento. Asimismo, luego de aplicar el preprocesamiento se obtuvo el 0% de inconsistencias, motivando con ello el incremento de los resultados del cálculo porcentual de la quiebra empresarial de 14.51% a 18.73%. Se obtuvo valores optimizados.

Tabla 11. Resultados antes y después del preprocesamiento

Matriz conjunto de datos	% Celdas Consistentes	% Inconsistencias	% Quiebra
Línea Base	72%	28%	14.51%
Preprocesado	100%	0%	18.73%
Nro Celdas	7,959		

Fuente. Elaboración propia.

4.1.2. Análisis de resultados de evaluación del performance de los algoritmos

4.1.2.1. Algoritmo XGBoost

XGBoost es un algoritmo mejorado que optimiza las actividades de clasificación y regresión. Asimismo, es importante el uso para la implementación de mejoras en el rendimiento de un conjunto de datos. Los autores Xiaolei et al. (2024) consideraron que el algoritmo es útil en la paralelización para mejorar el entrenamiento y optimizar los recursos en el manejo de valores faltantes que son útiles para la predicción.

En la Tabla 12 se puede apreciar la matriz de confusión generada a partir de los registros de las 379 empresas consultadas y que generaron un conjunto de datos. Asimismo, se identificó 335 empresas con predicción correcta y 44 empresas que

representaron a falsos positivos y negativos (fallos del modelo), de los cuales 30 empresas fueron falsos negativos que generan errores críticos a considerar.

Tabla 12. Matriz de confusión

Quiebra de empresas	Predicción			Total
	Positivo	Negativo	Porcentaje Correcto	
Positivo	111	30	29%	379
Negativo	14	224	59%	
	33%	67%	88%	

Fuente. Elaboración propia.

En la Tabla 13 se puede apreciar el despliegue del conjunto de datos de los registros de las 379 empresas con un algoritmo XGBoost, donde se obtuvo una precisión de 88%, exactitud de 86%, sensibilidad de 92% y f1-score de 90%. Los resultados obtenidos consideraron valores cercanos al 100%.

Tabla 13. Resultados XGBoost

	precision	sensibilidad	f1-score	total
	0.88	0.92	0.90	379
0	0.29	0.74	0.64	
1	0.59	0.18	0.26	
exactitud		0.86	379	
med-prom	0.75	0.74	0.41	379
pond-prom	0.79	0.80	0.72	379

Fuente. Elaboración propia.

En la Figura 4 se identificó la orientación de la curva ROC que fue generada por el algoritmo. Obtuvo el valor AUC=0.86.

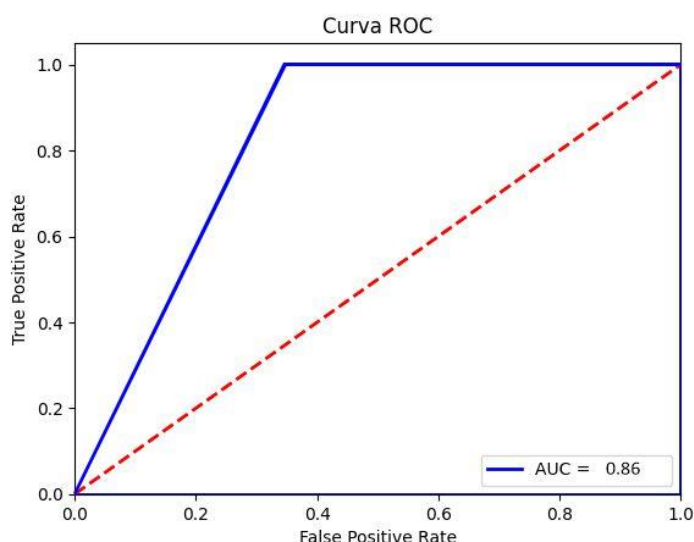


Figura 4. Curva ROC XGBoost

Fuente. Elaboración propia.

4.1.2.2. Algoritmo SVM

Los algoritmos SVM proporcionan una solución a problemas con valores atípicos que de diferentes grupos sustanciales. Los autores Yao et al. (2022) consideraron que el algoritmo proporciona resultados idóneos para decisiones empresariales mediante pronósticos y predicciones idóneas de riesgo que experimentan las empresas. En la Tabla 14 se puede apreciar la matriz de confusión generada a partir de los registros de las 379 empresas consultadas y que generaron un conjunto de datos. Asimismo, se identificó 300 empresas con predicción correcta y 79 empresas que representaron a falsos positivos y negativos (fallos del modelo), de los cuales 51 empresas fueron falsos negativos que generan errores críticos a considerar.

Tabla 14. Matriz de confusión

Quiebra de empresas	Predicción			Total
	Positivo	Negativo	Porcentaje Correcto	
Positivo	80	51	21%	379
Negativo	28	220	58%	
	28%	72%	79%	

Fuente. Elaboración propia.

En la Tabla 15 se puede apreciar el despliegue del conjunto de datos de los registros de las 379 empresas con un algoritmo SVM, donde se obtuvo una precisión de 79%, exactitud de 78%, sensibilidad de 85% y f1-score de 84%. Los resultados obtenidos consideraron valores cercanos al 100%.

Tabla 15. Resultados SVM

	precision	sensibilidad	f1-score	total
	0.79	0.85	0.84	379
0	0.21	0.69	0.57	
1	0.58	0.16	0.27	
exactitud		0.78	379	
med-prom	0.68	0.66	0.62	379
pond-prom	0.75	0.80	0.69	379

Fuente. Elaboración propia.

En la Figura 5 se identificó la orientación de la curva ROC que fue generada por el algoritmo. Obtuvo el valor AUC=0.78.

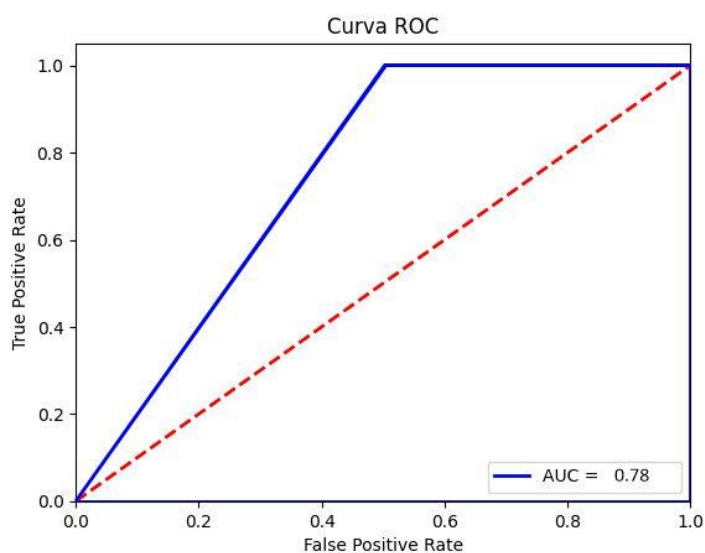


Figura 5. Curva ROC SVM

Fuente. Elaboración propia.

4.1.2.3. Algoritmo LSTM

El algoritmo LSTM trabaja con datos secuenciales y es útil para la predicción de los movimientos de series temporales que permitan recordar los datos relevantes en el tiempo sin limitaciones ni restricciones. Los autores Raza et al. (2022) consideraron que este algoritmo es aplicado en decisiones empresariales para identificar patrones y puntos focales que permitan generar una orientación hacia un problema específico. En la Tabla 16 se puede apreciar la matriz de confusión generada a partir de los registros de las 379 empresas consultadas y que generaron un conjunto de datos. Asimismo, se identificó 285 empresas con predicción correcta y 94 empresas que representaron a falsos positivos y negativos (fallos del modelo), de los cuales 31 empresas fueron falsos negativos que generan errores críticos a considerar.

Tabla 16. Matriz de confusión

Quiebra de empresas	Predicción			Total
	Positivo	Negativo	Porcentaje Correcto	
Positivo	60	31	16%	379
Negativo	63	225	59%	
	32%	68%	75%	

Fuente. Elaboración propia.

En la Tabla 17 se puede apreciar el despliegue del conjunto de datos de los registros de las 379 empresas con un algoritmo LSTM, donde se obtuvo una precisión de 75%, exactitud de 74%, sensibilidad de 82% y f1-score de 84%. Los resultados obtenidos consideraron valores cercanos al 100%.

Tabla 17. Resultados LSTM

	precision	sensibilidad	f1-score	total
	0.75	0.82	0.84	379
0	0.16	0.65	0.62	
1	0.59	0.17	0.22	

exactitud		0.74	379	
med-prom	0.59	0.58	0.62	379
pond-prom	0.62	0.70	0.71	379

Fuente. Elaboración propia.

En la Figura 6 se identificó la orientación de la curva ROC que fue generada por el algoritmo. Obtuvo el valor $AUC=0.74$.

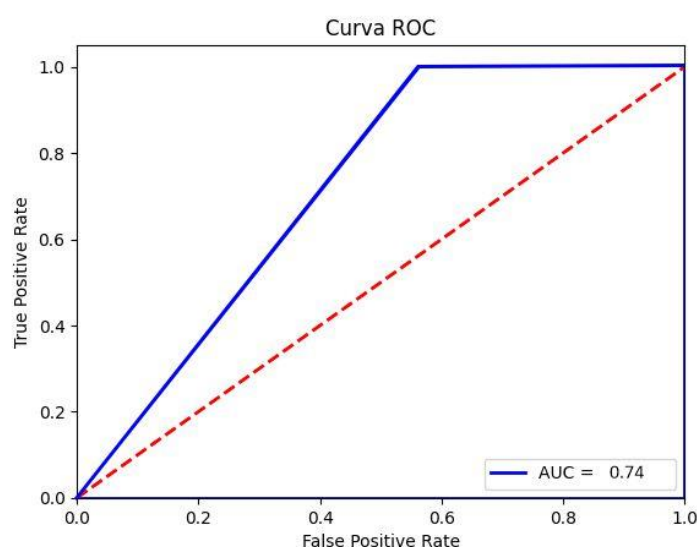


Figura 6. Curva ROC LSTM

Fuente. Elaboración propia.

4.1.2.4. Algoritmo de regresión logística

El algoritmo de regresión logística tiene técnicas utilizadas en empresas para soluciones de clasificación y regresión, además ayuda a propiciar la detección del fraude. Al respecto los autores Pendey et al. (2023) precisaron que el algoritmo de regresión logística permite identificar las relaciones entre dos variables y asociarlas al pronóstico empresarial. En la Tabla 18 se puede apreciar la matriz de confusión generada a partir de los registros de las 379 empresas consultadas y que generaron un conjunto de datos. Asimismo, se identificó 323 empresas con predicción correcta y 56 empresas que representaron a falsos positivos y negativos (fallos del modelo),

de los cuales 26 empresas fueron falsos negativos que generan errores críticos a considerar.

Tabla 18. Matriz de confusión de regresión logística

Quiebra de empresas	Predicción			Total
	Positivo	Negativo	Porcentaje Correcto	
Positivo	103	26	27%	379
Negativo	30	220	58%	
	35%	65%	85%	

Fuente. Elaboración propia.

En la Tabla 19 se puede apreciar el despliegue del conjunto de datos de los registros de las 379 empresas con un algoritmo de regresión logística, donde se obtuvo una precisión de 85%, exactitud de 83%, sensibilidad de 95% y f1-score de 90%. Los resultados obtenidos consideraron valores cercanos al 100%.

Tabla 19. Resultados del modelo de regresión logística

	precision	sensibilidad	f1-score	total
	0.85	0.95	0.90	379
0	0.27	0.73	0.58	
1	0.58	0.22	0.32	
exactitud		0.83	379	
med-prom	0.72	0.59	0.61	379
pond-prom	0.80	0.83	0.80	379

Fuente. Elaboración propia.

En la Figura 7 se identificó la orientación de la curva ROC que fue generada por el algoritmo. Obtuvo el valor AUC=0.83.

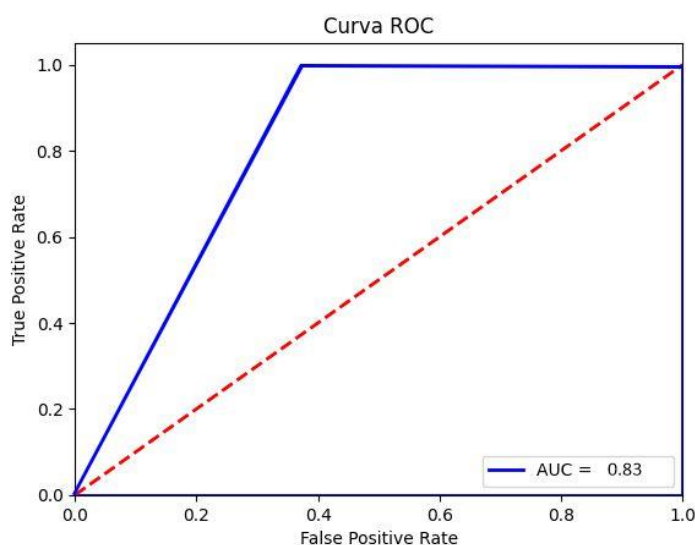


Figura 7. Curva ROC Regresión logística

Fuente. Elaboración propia.

4.1.2.5. Algoritmo de bosques aleatorios

El algoritmo de bosques aleatorios tiene técnicas que combinan múltiples árboles de decisión que se encuentran alineados para la obtención de un fin común. Asimismo, los autores Lee et al. (2019) consideraron que es importante el uso de este algoritmo debido a su facilidad de comprensión y alineamiento a los objetivos del estudio. En la Tabla 20 se puede apreciar la matriz de confusión generada a partir de los registros de las 379 empresas consultadas y que generaron un conjunto de datos. Asimismo, se identificó 316 empresas con predicción correcta y 63 empresas que representaron a falsos positivos y negativos (fallos del modelo), de los cuales 43 empresas fueron falsos negativos que generan errores críticos a considerar.

Tabla 20. Matriz de confusión de bosques aleatorios

Quiebra de empresas	Predicción			Total
	Positivo	Negativo	Porcentaje Correcto	
Positivo	88	43	23%	379
Negativo	20	228	60%	
	28%	72%	83%	

Fuente. Elaboración propia.

En la Tabla 21 se puede apreciar el despliegue del conjunto de datos de los registros de las 379 empresas con un algoritmo de bosques aleatorios. Asimismo, se procedió al despliegue del conjunto de datos, donde se obtuvo una precisión de 83%, exactitud de 83%, sensibilidad de 96% y f1-score de 90%.

Tabla 21. Resultados del modelo de bosques aleatorios

	precision	sensibilidad	f1-score	total
	0.83	0.96	0.90	379
0	0.23	0.90	0.74	
1	0.60	0.06	0.16	
exactitud		0.83	379	
med-prom	0.72	0.54	0.53	379
pond-prom	0.79	0.83	0.77	379

Fuente. Elaboración propia.

En la Figura 8 se identificó la orientación de la curva ROC que fue generada por el algoritmo. Obtuvo el valor AUC=0.83.

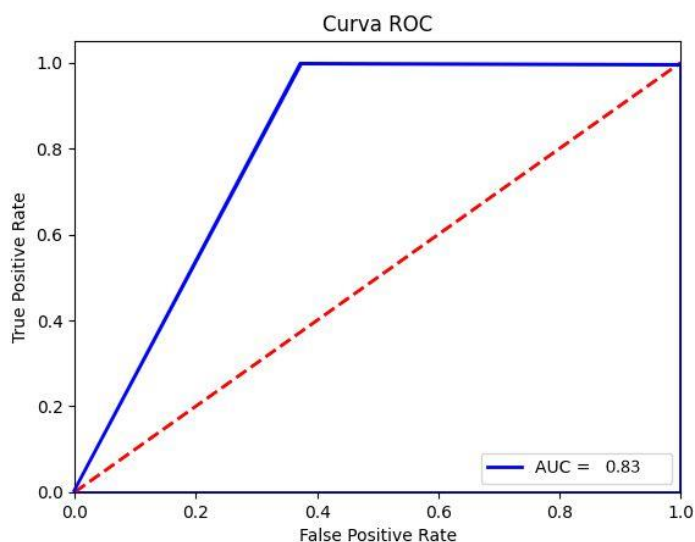


Figura 8. Curva ROC Bosques aleatorios

Fuente. Elaboración propia.

4.1.2.6. Algoritmo de red neuronal

Las redes neuronales se consideran como un algoritmo emergente que ha venido en aumento del área de aprendizaje automático. En ese sentido, los autores Raza et al. (2022) indicaron que las redes neuronales logran generar el procesamiento de una gran cantidad de variables y a una velocidad importante. Asimismo, utilizan técnicas útiles y novedosas para resolver problemas empresariales. En la Tabla 22 se puede apreciar la matriz de confusión generada a partir de los registros de las 379 empresas consultadas y que generaron un conjunto de datos. Asimismo, se identificó 341 empresas con predicción correcta y 38 empresas que representaron a falsos positivos y negativos (fallos del modelo), de los cuales 28 empresas fueron falsos negativos que generan errores críticos a considerar.

Tabla 22. Matriz de confusión de red neuronal

Quiebra de empresas	Predicción			Total
	Positivo	Negativo	Porcentaje Correcto	
Positivo	117	28	31%	379
Negativo	10	224	59%	
	34%	66%	90%	

Fuente. Elaboración propia.

En la Tabla 23 se puede apreciar el despliegue del conjunto de datos de los registros de las 379 empresas con un algoritmo de red neuronal. Se procedió al despliegue del conjunto de datos, donde se obtuvo una precisión de 90%, exactitud de 87%, sensibilidad de 97% y f1-score de 92%. Los resultados obtenidos consideraron valores cercanos al 100%. En ese sentido, se elige el algoritmo de red neuronal como base para la implementación del modelo optimizado para predecir la quiebra de empresas.

Tabla 23. Resultados del modelo de red neuronal

precision	sensibilidad	f1-score	total
0.90	0.97	0.92	379

	0	0.31	0.54	0.43	
	1	0.59	0.43	0.49	
exactitud			0.87	379	
med-prom		0.74	0.69	0.71	379
pond-prom		0.85	0.87	0.86	379

Fuente. Elaboración propia.

En la Figura 9 se identificó la orientación de la curva ROC que fue generada por el algoritmo. Obtuvo el valor AUC=0.87.

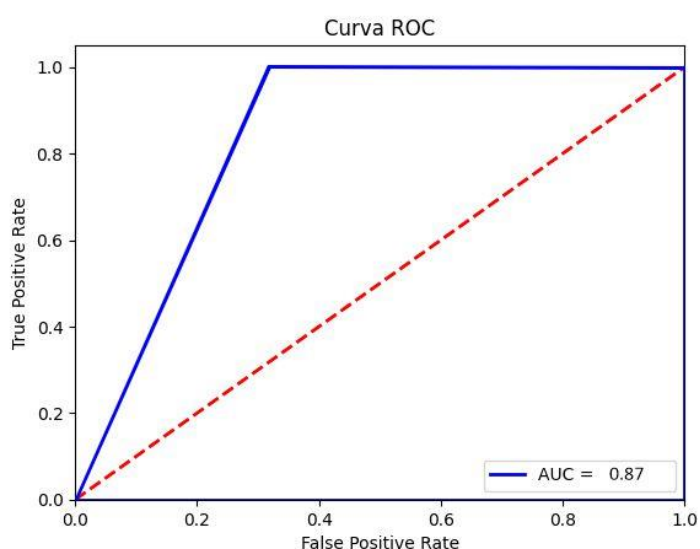


Figura 9. Curva ROC Red neuronal

Fuente. Elaboración propia.

El algoritmo fue seleccionado bajo los criterios indicados en la etapa 3.9.3.2. Evaluación de algoritmos. Se identificaron los siguientes algoritmos: XGBoost, SVM, LSTM, regresión logística, bosques aleatorios y redes neuronales. Asimismo, para obtener el algoritmo idóneo se consideraron los mayores resultados porcentuales en las métricas de precisión, exactitud, sensibilidad y f1-score. Adicionalmente a ello, se compararon los resultados obtenidos de la curva ROC, los cuales se precisaron en la Figura 10. Se logró identificar que los valores obtenidos del AUC fueron determinantes para establecer un criterio sobre los algoritmos que tienen un mejor comportamiento y rendimiento para el conjunto de

datos objeto de estudio. El área más amplia con valores cercanos al 1 fueron los de mejor propuesta.

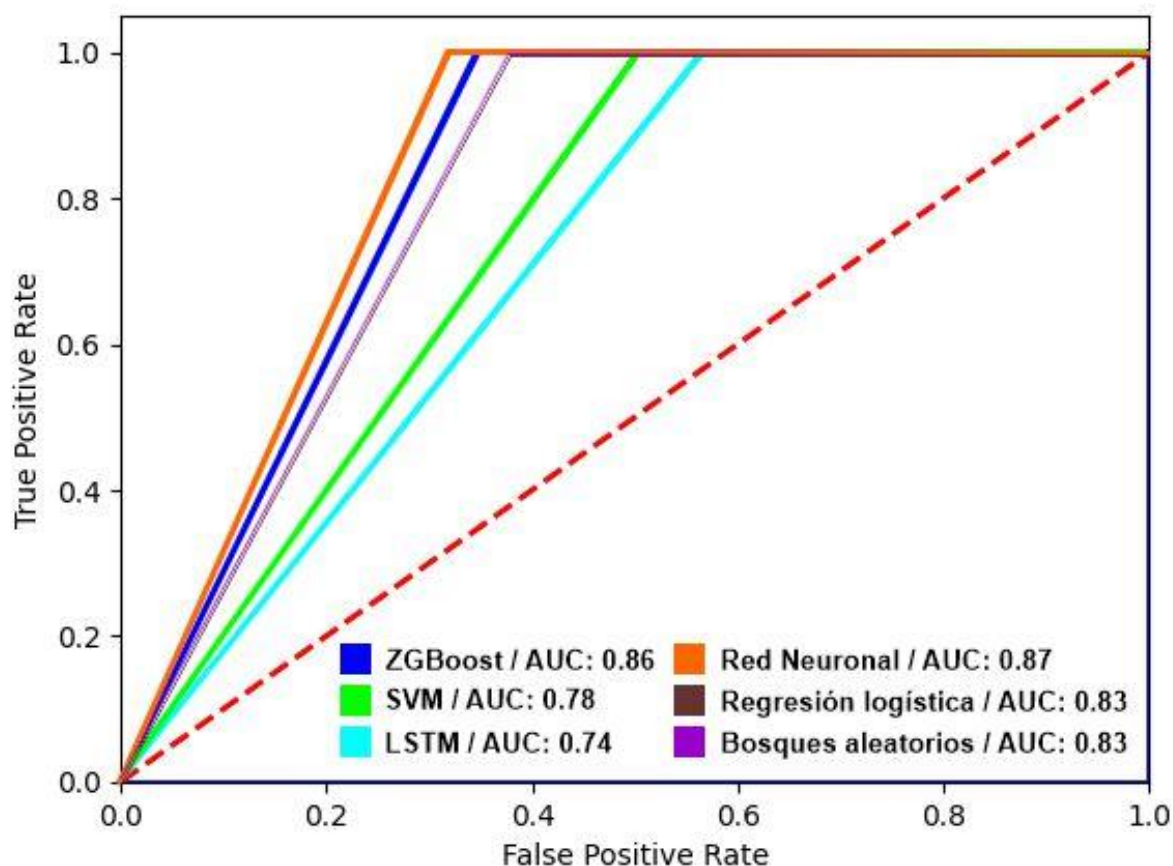


Figura 10. Curva ROC Algoritmos

Fuente. Elaboración propia.

4.1.3. Análisis de la correlación de factores relevantes de las empresas

4.1.3.1. Análisis e identificación de variables relevantes

Se utilizó el análisis de datos en distribuciones y gráficos de líneas de las variables: Quiebra, Meses_A, Capital, Planes, SI, Calificación, Tasa_utilidades, Satisfacción_E, Acceso_Credito, Acceso_mercados, Personal, Letras_Por_Pagar, Sociedades, Rendimiento_Pago, Competidores_D, Imprevistos, Siniestros, Tasa_utilidades, Rentabilidad, Eficiencia e Insolvencia, donde se identificó niveles

de predominio que requieren correlacionar hacia cada factor resultante. Asimismo, se identificaron variables del conjunto de datos que se correlacionaron entre ellas, las cuales dependían del contexto propio de la empresa.

Los autores Alsheret et al. (2022) consideraron que los gráficos que proporciona la biblioteca Seaborn (Python) son ideales para establecer correlaciones entre variables de un conjunto de datos. Asimismo, precisaron que esta herramienta analiza mediante patrones y relaciones entre datos que obtienen puntuaciones representativas y según ponderaciones realizadas. Para la obtención de los valores de correlación se utilizó la siguiente fórmula:

$$r = \frac{\sum (x - \bar{x}) (y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Donde r representa la correlación entre x, y (medidas), $\bar{x}$, $\bar{y}$ (media de las medidas). Asimismo, para encontrar la correlación se realizó mediante el cálculo de la multiplicación de las diferencias de las variables x, $\bar{x}$, y, $\bar{y}$. Luego de ello, se divide por el resultado de la raíz cuadrada de la multiplicación de la suma de los cuadrados de las diferencias entre x, $\bar{x}$; y la suma de los cuadros de la diferencia de y, $\bar{y}$. En la Figura 11 se aprecia que se obtuvo valores de correlación de 0 a 100%.

NRO	MATRIZ CORRELACIÓN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Quiebra		55%	45%	35%	12%	18%	40%	30%	46%	51%	40%	45%	45%	30%	35%	30%	55%	30%	34%	46%	40%
2	Meses_A	55%		48%	34%	15%	40%	45%	85%	66%	55%	30%	62%	46%	30%	55%	30%	34%	46%	93%	30%	96%
3	Capital	45%	48%		30%	13%	34%	46%	40%	32%	45%	35%	30%	30%	90%	40%	50%	12%	35%	55%	50%	53%
4	Planes	35%	34%	30%		20%	25%	28%	45%	30%	34%	46%	40%	32%	20%	35%	30%	45%	35%	36%	30%	34%
5	SI	12%	15%	13%	20%		23%	24%	19%	18%	45%	12%	10%	55%	35%	34%	55%	40%	32%	10%	15%	12%
6	Calificación	18%	40%	34%	25%	23%		45%	40%	30%	35%	50%	50%	52%	55%	46%	55%	38%	50%	50%	44%	42%
7	Capacidad_E	40%	45%	46%	28%	24%	45%		30%	25%	20%	52%	45%	55%	50%	55%	55%	25%	20%	15%	14%	12%
8	Satisfacción_E	30%	85%	40%	45%	19%	40%	30%		10%	15%	20%	42%	55%	40%	55%	35%	45%	55%	15%	20%	42%
9	Acceso_Credito	46%	66%	32%	30%	18%	30%	25%	10%		55%	30%	45%	55%	30%	45%	30%	55%	30%	34%	46%	30%
10	Acceso_Mercados	51%	55%	45%	34%	45%	35%	20%	15%	55%		55%	30%	45%	30%	45%	18%	40%	30%	46%	55%	25%
11	Personal	40%	30%	35%	46%	12%	50%	52%	20%	30%	55%		30%	55%	30%	45%	45%	55%	50%	55%	55%	30%
12	Letras_por_pagar	45%	62%	30%	40%	10%	50%	45%	42%	45%	30%	30%		55%	30%	45%	42%	55%	40%	55%	35%	36%
13	Sociedades	45%	46%	30%	32%	55%	52%	55%	55%	55%	45%	55%	55%		30%	46%	55%	38%	50%	20%	42%	43%
14	Rendimiento_Pago	30%	30%	90%	20%	35%	55%	50%	40%	30%	30%	30%	30%	30%		38%	34%	46%	40%	32%	20%	45%
15	Competidores_D	35%	55%	40%	35%	34%	46%	55%	55%	45%	45%	45%	45%	46%	38%		46%	40%	32%	46%	40%	32%
16	Imprevistos	30%	30%	50%	30%	55%	55%	55%	35%	30%	18%	45%	42%	55%	34%	46%		46%	46%	46%	40%	32%
17	Siniestros	55%	34%	12%	45%	40%	38%	25%	45%	55%	40%	55%	38%	46%	40%	46%	46%		30%	30%	34%	46%
18	Tasa_utilidades	30%	46%	35%	35%	32%	50%	20%	55%	30%	30%	50%	40%	50%	40%	32%	46%	30%		30%	34%	46%
19	Rentabilidad	34%	93%	55%	36%	10%	50%	15%	15%	34%	46%	55%	55%	20%	32%	46%	46%	30%	30%		34%	46%
20	Eficiencia	46%	30%	50%	30%	15%	44%	14%	20%	46%	55%	55%	35%	42%	20%	40%	40%	34%	34%	34%		20%
21	Insolvencia	40%	96%	53%	34%	12%	42%	12%	42%	30%	25%	30%	36%	43%	45%	32%	32%	46%	46%	46%	20%	

Figura 11. Las correlaciones entre las variables

Fuente. Elaboración propia.

Se consideraron variables relevantes con valores más cercanos al 100% y que se relacionaron con el contexto propio de la empresa: Meses_A, Capital, Satisfacción_E, Rendimiento_Pago, Rentabilidad e Insolvencia. Adicionalmente a ello, se consideraron variables que se encontraron entre el 60% al 80% como variables de complemento: Acceso_Crédito y Letras_por_pagar.

En la Tabla 24 se consideraron las ocho (8) variables requeridas para iniciar el entrenamiento y limpieza de los datos. Las seis (6) variables relevantes fueron aquellas que permitieron conseguir la optimización del modelo y las dos (2) variables complementarias fueron las que permitieron establecer criterios de mejora.

Tabla 24. Variables relevantes

Variables	Valor de correlación	Variable Correlacionada
Meses_A	85%	Satisfacción_E
	66%	Acceso_Credito
	62%	Letras_por_pagar
	93%	Rentabilidad
	96%	Insolvencia
Capital	90%	Rendimiento_Pago
Satisfacción_E	85%	Meses_A
Acceso_Credito	66%	Meses_A
Letras_por_pagar	62%	Meses_A
Rendimiento_Pago	90%	Capital
Rentabilidad	93%	Meses_A
Insolvencia	96%	Meses_A

Fuente. Elaboración propia.

Se realizaron las técnicas de Shap y Lime para identificar los factores que tienen mayor relevancia para la predicción y de esta manera establecer una comparación con los factores obtenidos en la técnica de correlación. Para la técnica Shap se realizó el despliegue de conjunto de datos para conocer los factores más importantes del modelo. En ese sentido, se obtuvo el siguiente diagnóstico presentado en la Figura 12:

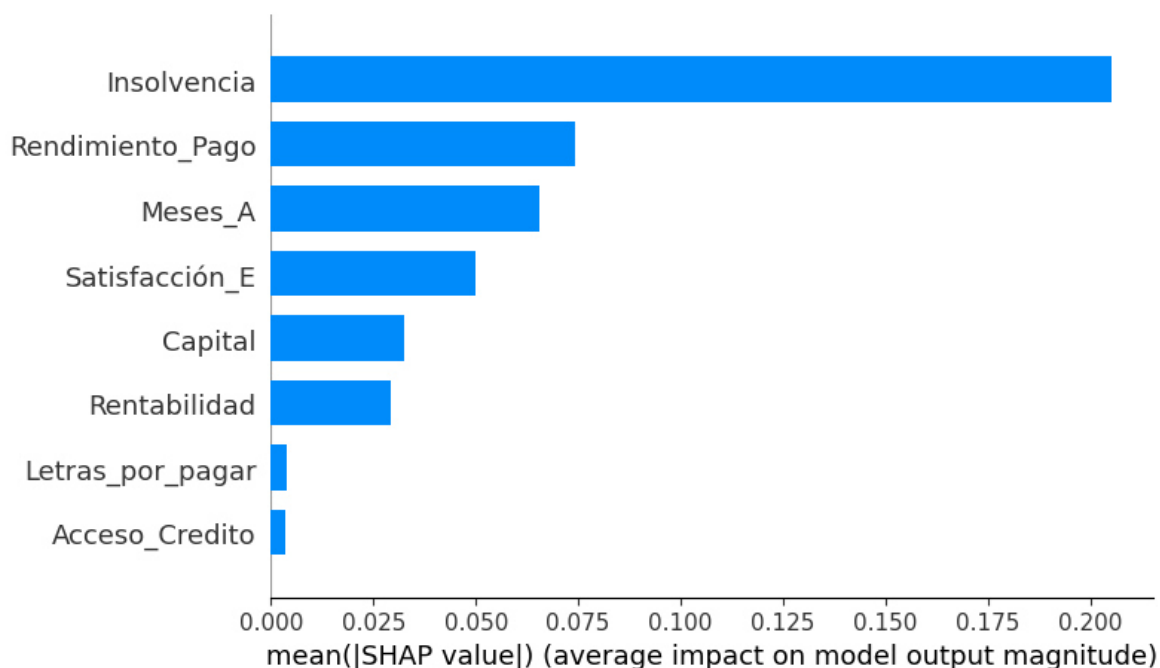


Figura 12. Factores con mayor incidencia

Fuente. Elaboración propia.

La técnica Shap permitió identificar que el factor Insolvencia es el factor más relevante del modelo propuesto según los resultados del conjunto de datos. Asimismo, se identificaron otros factores que brindan relevancia como: Rendimiento_Pago, Meses_A, Satisfacción_E, Capital y Rentabilidad. Los factores Letras_por_pagar y Acceso_Credito fueron identificados como factores que brindan un complemento al modelo.

En la Figura 13 se puede apreciar el impacto (alto o bajo) de cada factor con la matriz del conjunto de datos, donde se obtuvo que el factor Insolvencia, Meses_A y Satisfacción_E fueron los factores que presentaron una predicción mayor respecto al impacto para el modelo.

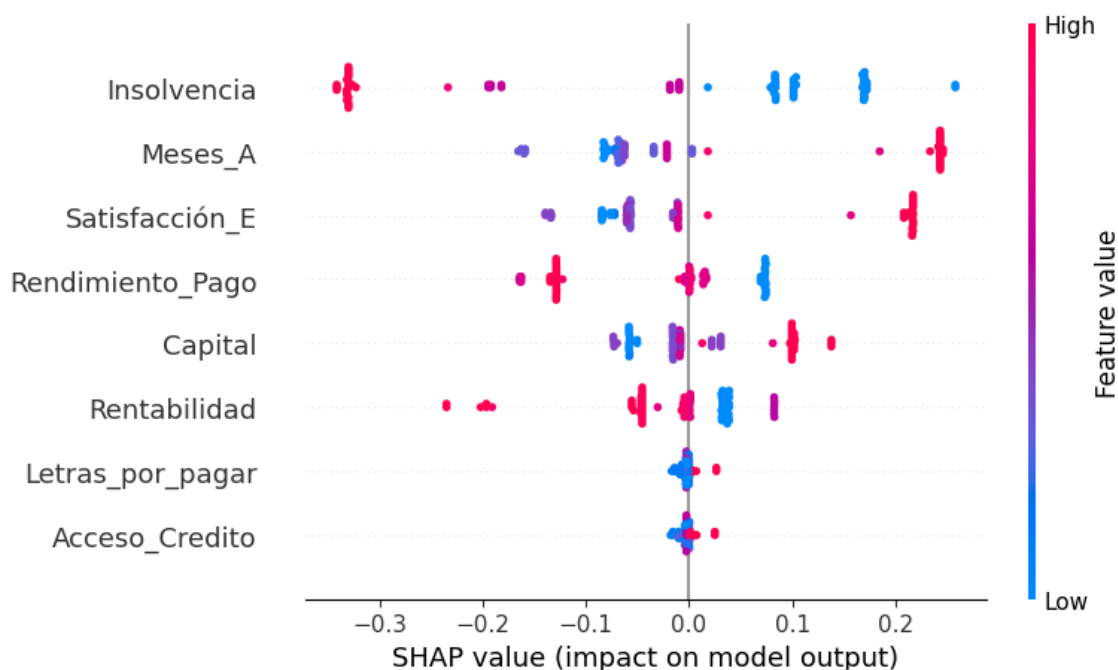


Figura 13. Impacto de factores

Fuente. Elaboración propia.

Respecto a la técnica Lime se realizó para identificar las características que más influyen con los factores relevantes del modelo. De esta manera, se generó la interpretación más eficiente de los resultados obtenidos del modelo. En la Figura 14 se precisaron los resultados luego del despliegue de la técnica.

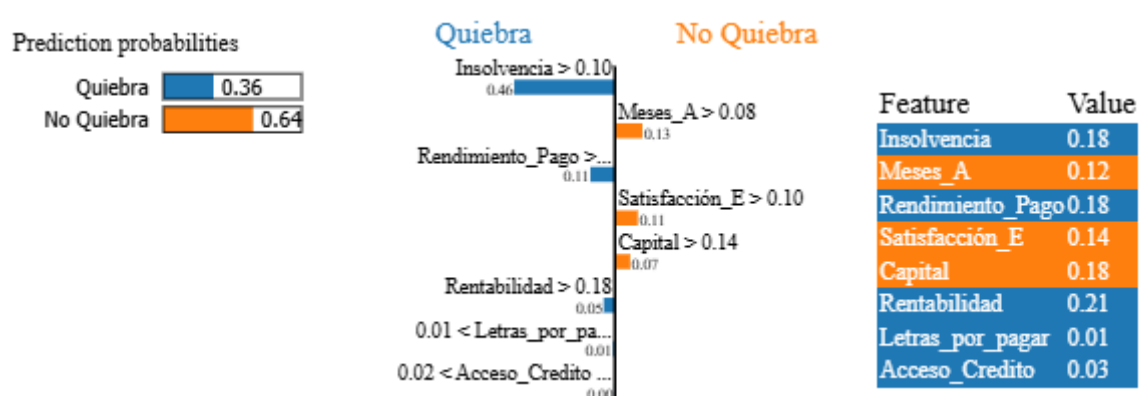


Figura 14. Características influyentes

Fuente. Elaboración propia.

Los factores que más influyeron en el modelo fueron precisados en la Figura 15, donde se obtuvo: Insolvencia (mayor a 0.10 con un 46% para la quiebra), Meses_A (mayor a 0.08 con un 13% a la no quiebra), Rendimiento_Pago (mayor a 0.14 con un 11% de quiebra), Satisfacción_E (mayor a 0.10 con un 11% a la no quiebra), Capital (mayor a 0.14 con un 7% a la no quiebra) y Rentabilidad (mayor a 0.18 con un 5% de quiebra). Asimismo, los factores que brindan un complemento fueron: Letras_por_pagar (mayor a 0.01 y menor o igual a 0.01 con un 1% de quiebra) y Acceso_Credito (mayor a 0.02 y menor o igual a 0.04 menor a 1%). Asimismo, según los resultados luego de aplicar de la técnica en la matriz del conjunto datos, se obtuvo que las predicciones indicaron la presencia de Quiebra empresarial en 36% de empresas y no quiebra en 64%.

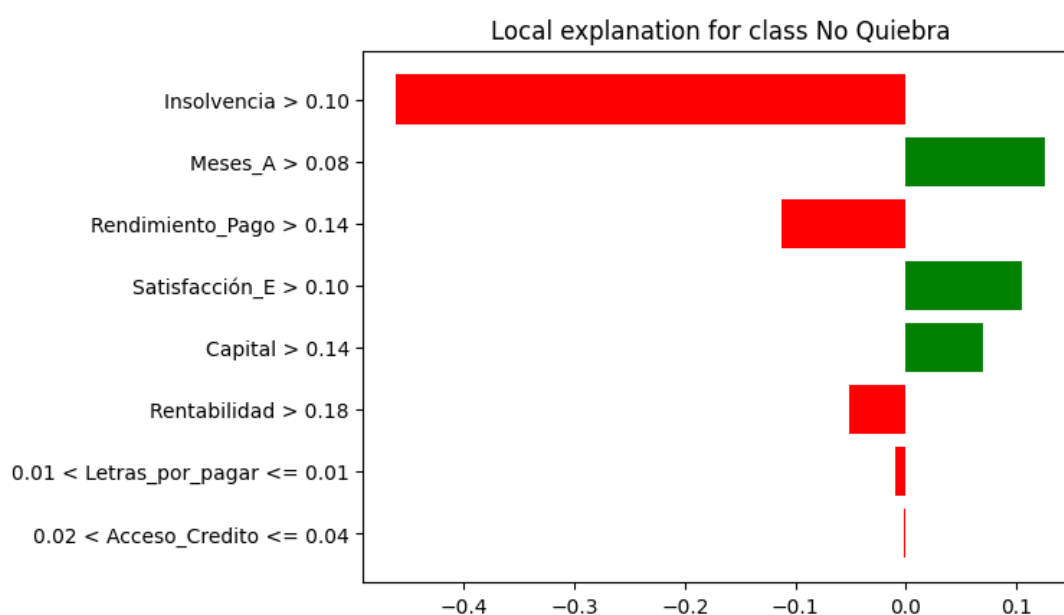


Figura 15. Despliegue de características influyentes

Fuente. Elaboración propia.

Los resultados obtenidos luego de la aplicación de la técnica PCA para la reducción de la dimensionalidad permitieron identificar las variables que tienen mayor incidencia para el tratamiento del conjunto de datos, se determinó que diez (10) variables son las más representativas y que permitieron reducir el tamaño del conjunto de datos, definiéndose solo con los datos registrados de las variables:

Meses_A, Capital, Satisfacción_E, Acceso_Credito, Letras_por_pagar, Rendimiento_Pago, Rentabilidad, Insolvencia, Calificación y Tasa_utilidades. En la Tabla 25 se precisaron los datos obtenidos.

Tabla 25. Resultados PCA

Variables	Datos
Meses_A	379
Capital	379
Satisfacción_E	379
Acceso_Credito	379
Letras_por_pagar	379
Rendimiento_Pago	379
Rentabilidad	379
Insolvencia	379
Calificación	379
Tasa_utilidades	379
%Consistencia	100%
Dimen	10
Total	3790
%Red Dimen	47.62%

Fuente. Elaboración propia.

Se identificaron que las variables obtenidas de PCA fueron comunes a los factores relevantes obtenidos con las técnicas de correlación, Shap y Lime; en ese sentido, se confirmó la similitud de resultados, la equivalencia entre las operaciones realizadas y el uso de los factores relevantes para todo el estudio. Posterior a ello, se procedió a construir nuevas características a partir de la correlación de variables, donde se generaron los siguientes índices:

$$\text{Índice de riesgo} = \frac{\text{Insolvencia}}{\text{Meses_A}} \%$$

Donde los datos obtenidos corresponden al cálculo realizado entre las variables Insolvencia y Meses_A, los resultados se consideraron como el porcentaje del riesgo de préstamo que presenta para una Institución financiera, cuanto más insolvencia presenta en el tiempo de operaciones mayor será el porcentaje.

$$\text{Indice Confiabilidad} = \frac{(\text{Rendimiento_Pago} * \text{Meses_A})}{100} \%$$

Donde los datos obtenidos corresponden al cálculo realizado entre las variables Rendimiento_Pago y Meses_A, los resultados se consideraron como el porcentaje de confianza que tiene una empresa frente a un potencial préstamo de una Institución financiera, cuanto más rendimiento de pago presenta en el tiempo de operaciones mayor será el porcentaje de confianza.

$$\text{Indice Sostenibilidad} = \frac{(\text{Calificación} * \text{Meses_A})}{100} \%$$

Donde los datos obtenidos corresponden al cálculo realizado entre las variables Calificación y Meses_A, los resultados se consideraron como el porcentaje de sostenerse en las operaciones empresariales y que pueden motivar a un préstamo, cuanto más calificación presenta en el tiempo de operaciones mayor será el porcentaje de sostenibilidad y por tanto obtendrá mejores condiciones.

$$\text{Indice Solidez} = \frac{(\text{Tasa_utilidades} * \text{Meses_A})}{100}$$

Donde los datos obtenidos corresponden al cálculo realizado entre las variables Tasa_utilidades y Meses_A, los resultados se consideraron como el valor que brinda solidez a una empresa y que hace atractivo a potenciales préstamos de una Institución financiera, cuantas más utilidades presenta en el tiempo de operaciones mayor será el valor de solidez empresarial.

4.1.3.2. Análisis de resultados de la implementación

El algoritmo seleccionado consideró 21 variables que tienen características propias del empleado según el contexto interno de la empresa. Inicialmente se tiene una precisión de 90%, exactitud de 87%, sensibilidad de 95% y f1-score de 92%. Asimismo, el modelo propuesto por los autores Alsheref et al. (2022) fue base para la complementación hacia el modelo optimizado que se propuso en la presente investigación: Modelo optimizado de aprendizaje automático (MO2A).

4.1.3.3. Análisis de resultados del entrenamiento

En la Figura 16 se precisó el entrenamiento con 100 eventos que fueron necesarios para mejorar las métricas mediante un proceso de aprendizaje recurrente. Asimismo, para generar la sentencia de código se necesitó de las librerías tensorflow y keras, los cuales son ideales en la construcción de modelos de aprendizaje profundo. La sentencia que se ejecutó fue:

```
import tensorflow / import keras
model = tf.keras.models.Sequential()
model.compile(optimizer, loss, metrics=['accuracy'])
epochs_hist = model.fit(X_train, y_train, epochs = 100, batch_size=10)
epochs_hist.history.keys()
plt.plot(epochs_hist.history['accuracy'])
plt.title("Tasa de Acierto del Modelo durante el Entrenamiento")
plt.xlabel("Epochs")
plt.ylabel("Accuracy de Entrenamiento")
plt.legend(["Accuracy de Entrenamiento"])
```

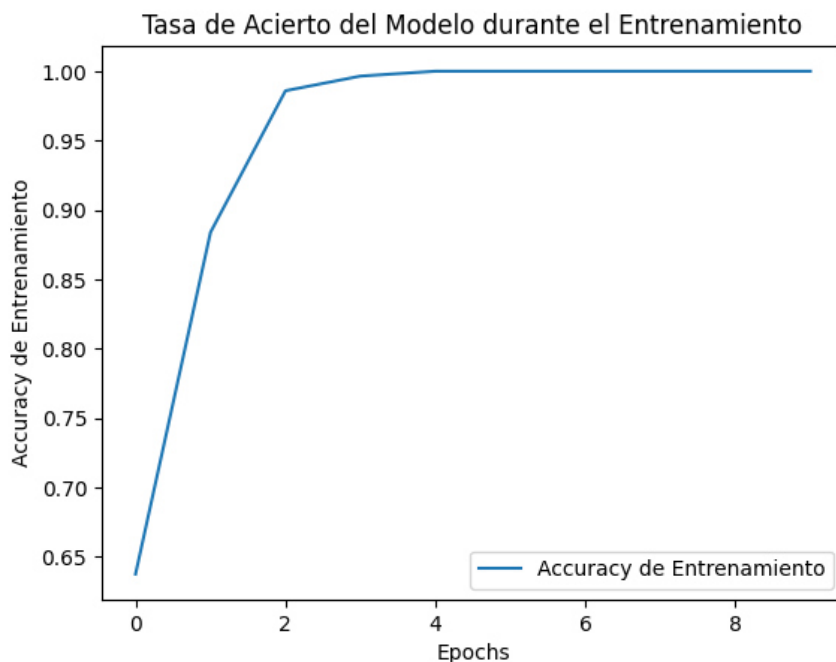


Figura 16. Tasa de acierto del modelo durante el entrenamiento
Fuente. Elaboración propia.

4.1.3.4. Factores para la quiebra de empresas

Se obtuvo resultados sobre la variable relevante de Rendimiento_Pago, donde se identificó que se presenta una cantidad mayor de empresas que quiebran al tener un menor rendimiento de pago de sus obligaciones y las empresas que no quiebran son aquellas que en su mayoría se encuentran con mayor rendimiento de pago. Adicionalmente a ello, se identificó que una cantidad menor de empresas que quiebran son los que se no podrán recuperar a futuro por encontrarse con deudas impagables. Asimismo, con la sentencia kdeplot de la librería Searborn en Python (técnica de visualización) se pudo identificar los rangos altos y bajos de la variable. La sentencia que se ejecutó con kdeplot fue:

```
plt.figure(figsize=(12, 7))
sns.kdeplot(left_df['Rendimiento_Pago'], label = "Rendimiento Pago Bajo",
            color = 'r', fill=True)
sns.kdeplot(stayed_df['Rendimiento_Pago'], label = "Rendimiento Pago Alto",
            color = 'b', fill=True)
plt.xlabel('Rendimiento_Pago')
```

En la Figura 17 se precisaron los resultados de la orientación del conjunto de datos.

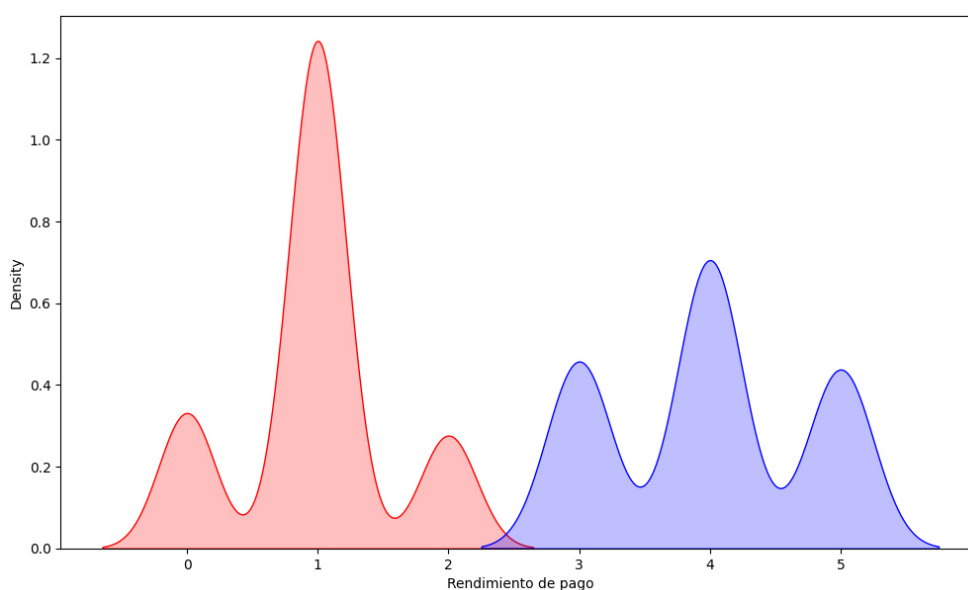


Figura 17. Resultados factor rendimiento de pago

Fuente. Elaboración propia.

Se obtuvo resultados sobre la variable relevante de Satisfacción_E, donde se identificó que se presenta una cantidad mayor de empresas que quiebran al tener menor satisfacción empresarial (calificación de 0) y las empresas que no quiebran son en su mayoría las que se encuentran con mayor satisfacción empresarial (4). Asimismo, con la sentencia kdeplot de la librería Searborn en Python (técnica de visualización) se pudo identificar los rangos altos y bajos de la variable.

La sentencia que se ejecutó con kdeplot fue:

```
plt.figure(figsize=(12, 7))
sns.kdeplot(left_df['Satisfaccion_E'],label="Satisfacción Bajo",color = 'r', fill=True)
sns.kdeplot(stayed_df['Satisfaccion_E'],label="Satisfacción Alto",color='b',fill=True)
plt.xlabel('Satisfaccion_E')
```

En la Figura 18 se precisaron los resultados de la orientación del conjunto de datos.

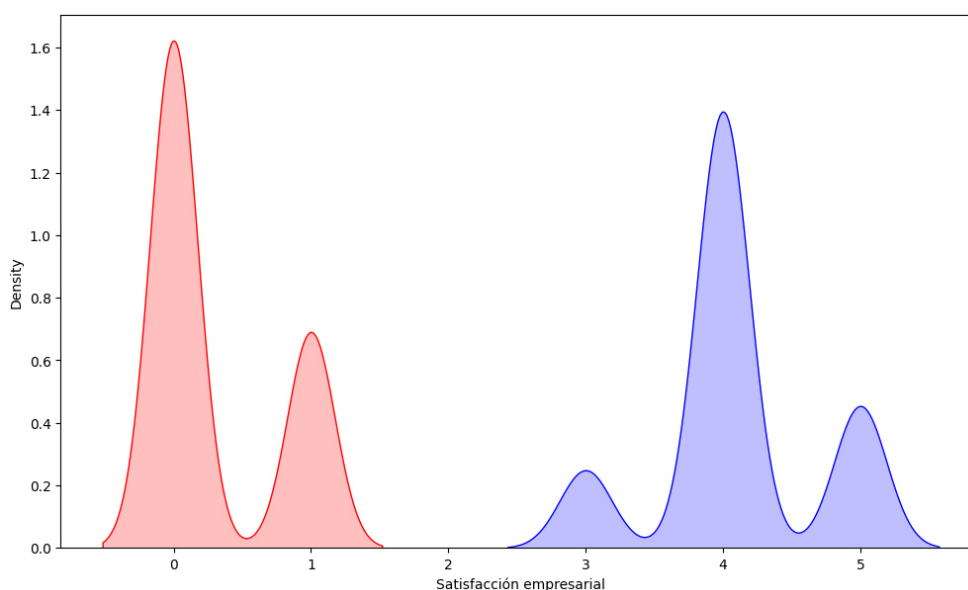


Figura 18. Resultados factor satisfacción empresarial

Fuente. Elaboración propia.

Se obtuvo resultados sobre la variable relevante de Capital, donde se identificó que se presenta una cantidad mayor de empresas que quiebran al tener menor capital (menores de S/ 5,000.00) y las empresas que no quiebran son en su mayoría las que se encuentran con mayor capital (mayores a S/ 5,000.00). Asimismo, con la

sentencia kdeplot de la librería Searborn en Python (técnica de visualización) se pudo identificar los rangos altos y bajos de la variable.

La sentencia que se ejecutó con kdeplot fue:

```
plt.figure(figsize=(12, 7))
sns.kdeplot(left_df['Capital'], label = "Capital Bajo",color = 'r', fill=True)
sns.kdeplot(stayed_df['Capital'], label = "Capital Alto",color = 'b', fill=True)
plt.xlabel('Capital')
```

En la Figura 19 se precisaron los resultados de la orientación del conjunto de datos.

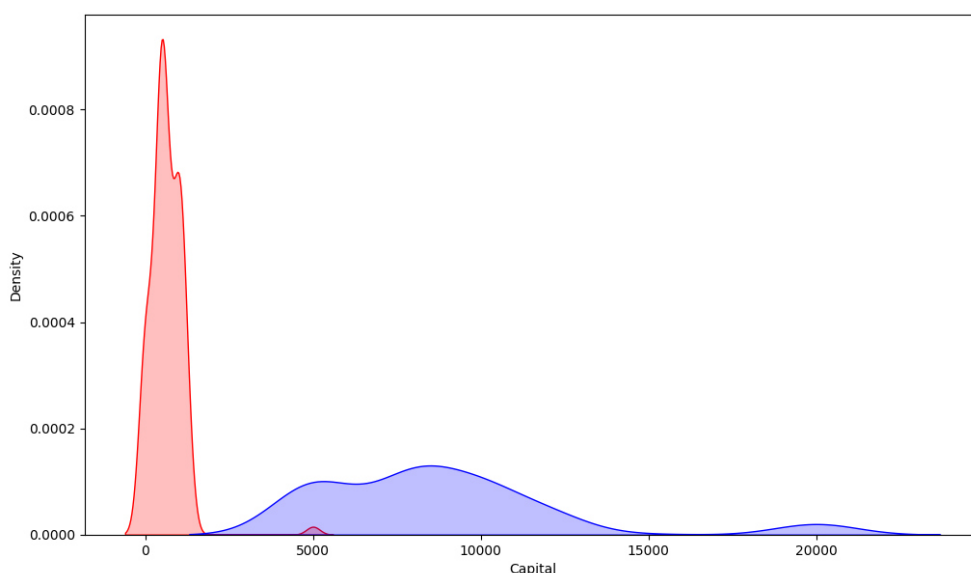


Figura 19. Resultados factor capital

Fuente. Elaboración propia.

Se obtuvo resultados sobre la variable relevante de Insolvencia, donde se identificó que se presenta una cantidad mayor de empresas que quiebran al tener insolvencia alta. Asimismo, se identificaron empresas que tienen insolvencia baja por la que tienen mayor capacidad de sostenerse con el tiempo. Con la sentencia kdeplot de la librería Searborn en Python (técnica de visualización) se pudo identificar los rangos altos y bajos de la variable.

La sentencia que se ejecutó con kdeplot fue:

```
plt.figure(figsize=(12, 7))
sns.kdeplot(left_df['Insolvencia'], label = "Insolvencia Bajo",
            color = 'r', fill=True)
sns.kdeplot(stayed_df['Insolvencia'], label = "Insolvencia Alto",
            color = 'b', fill=True)
plt.xlabel('Insolvencia')
```

En la Figura 20 se precisaron los resultados de la orientación del conjunto de datos.

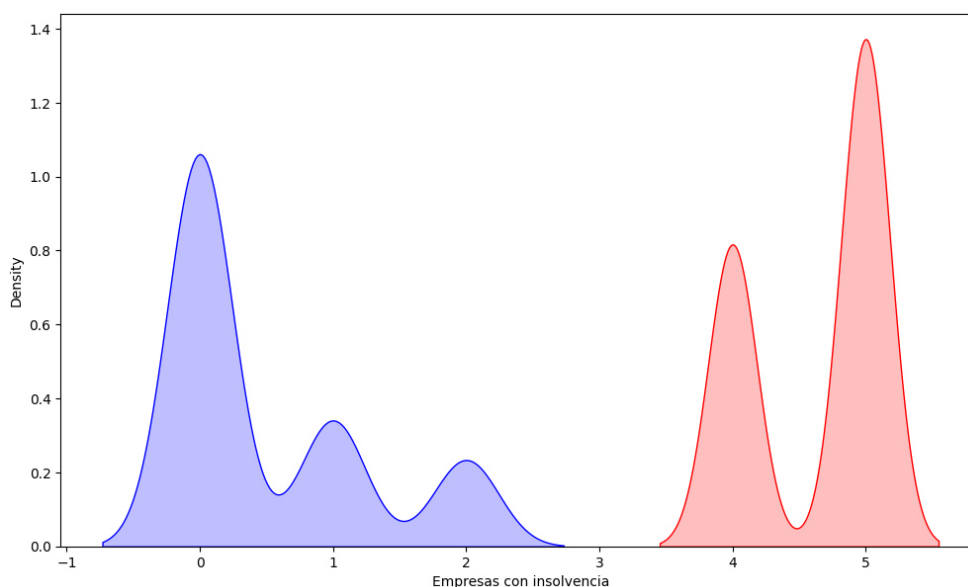


Figura 20. Resultados factor insolvencia

Fuente. Elaboración propia.

Se obtuvo resultados sobre la variable relevante de tiempo de actividad de la empresa (Meses_A), donde se identificó que una cantidad menor de tiempo de actividad hace que la empresa se encuentre en mayor riesgo de quiebra. Asimismo, con la sentencia kdeplot de la librería Seaborn en Python (técnica de visualización) se pudo identificar los rangos altos y bajos de la variable.

La sentencia que se ejecutó con kdeplot fue:

```
plt.figure(figsize=(12, 7))
```

```

sns.kdeplot(left_df['Meses_A'], label = "Tiempo de actividad Bajo",
            color = 'r', fill=True)
sns.kdeplot(stayed_df['Meses_A'], label = "Tiempo de actividad Alto",
            color = 'b', fill=True)
plt.xlabel('Meses_A')

```

En la Figura 21 se precisaron los resultados de la orientación del conjunto de datos.

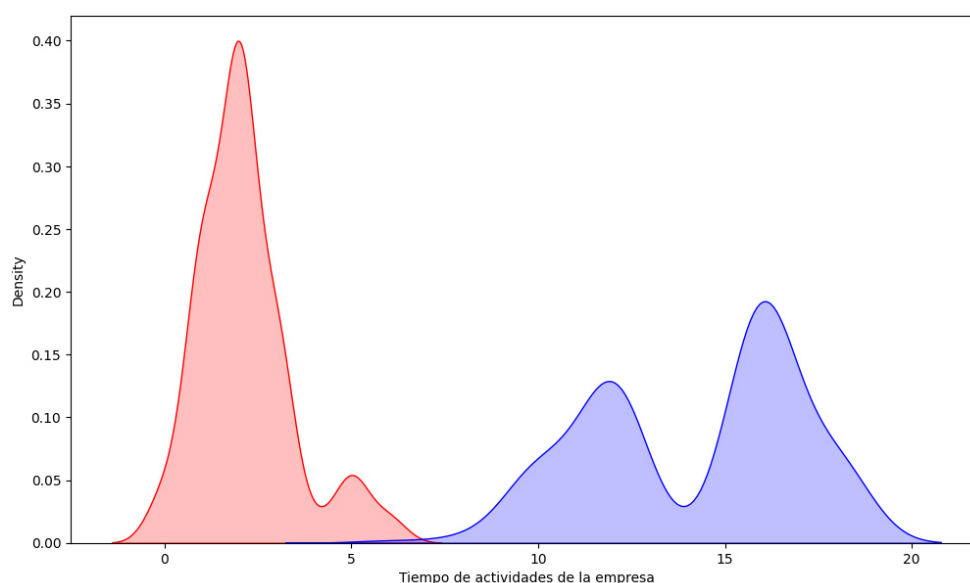


Figura 21. Resultados factor tiempo de actividad
Fuente. Elaboración propia.

Se obtuvo resultados sobre la variable relevante de Rentabilidad, donde se identificó que se presenta una cantidad mayor de empresas que quiebran al tener menor Rentabilidad (calificación de 0) y las empresas que no quiebran son en su mayoría las que se encuentran con mayor Rentabilidad (5). Asimismo, con la sentencia `kdeplot` de la librería `Searborn` en `Python` (técnica de visualización) se pudo identificar los rangos altos y bajos de la variable.

La sentencia que se ejecutó con `kdeplot` fue:

```

plt.figure(figsize=(12, 7))
sns.kdeplot(left_df['Rentabilidad'], label="Rentabilidad Baja", color = 'r', fill=True)

```

```
sns.kdeplot(stayed_df['Rentabilidad'], label="Rentabilidad Alta",color = 'b', fill=True)
plt.xlabel('Rentabilidad')
```

En la Figura 22 se precisaron los resultados de la orientación del conjunto de datos.

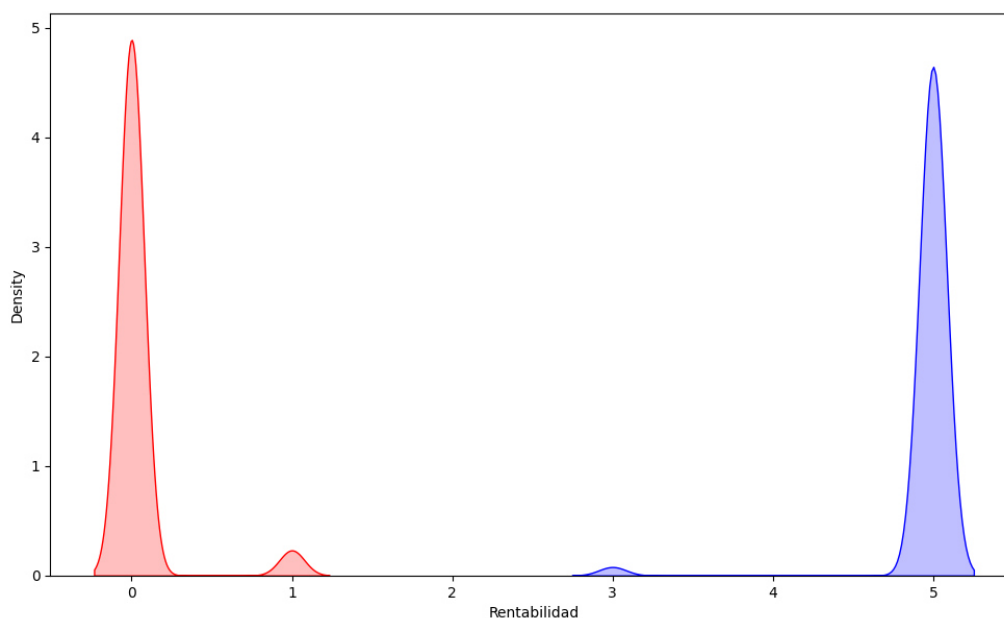


Figura 22. Resultados factor rentabilidad

Fuente. Elaboración propia.

4.1.3.5. Correlaciones entre variables

Las correlaciones significan la identificación de aspectos que se generan a partir de los atributos de cada variable y que en conjunto significan un factor relevante. Se determinó las siguientes correlaciones: Rendimiento de pago con capital, Rentabilidad, Satisfacción Empresarial e Insolvencia con Tiempo de actividad (Meses_A).

El rendimiento de pago es una variable importante en todo modelo de predicción puesto que con ello se puede identificar si su tratamiento es determinante para la quiebra empresarial. Asimismo, el rendimiento de pago mayor tiene un grado de identificación muy alto dentro de la variable relevante para este modelo optimizado. Lo contrario al rendimiento bajo que es un indicador de posible quiebra empresarial.

En la Tabla 26 se precisaron resultados que fueron generados sobre el rendimiento de pago mediante la estadística descriptiva.

Tabla 26. Tabla de frecuencias de la variable rendimiento de pago

Rendimiento_	Pago	F	P%	P% válido	P% acumulado
Válido	0	24	6,3	6,3	6,3
	1	90	23,7	23,7	30,1
	2	20	5,3	5,3	35,4
	3	70	18,5	18,5	53,8
	4	108	28,5	28,5	82,3
	5	67	17,7	17,7	100,0
	Total	379	100,0	100,0	

Fuente. Elaboración propia.

En la Tabla 27 se precisaron resultados que fueron generados sobre el capital mediante la estadística descriptiva, los cuales fueron:

Tabla 27. Tabla de frecuencias de la variable capital

Capital	F	P%	P% válido	P% acumulado
Válido	0	22	5,8	5,8
	200	5	1,3	7,1
	500	57	15,0	22,2
	800	5	1,3	23,5
	1000	34	9,0	32,5
	1200	10	2,6	35,1
	5000	70	18,5	53,6
	6000	2	,5	54,1
	7000	1	,3	54,4
	7500	1	,3	54,6
	8000	75	19,8	74,4
	10000	55	14,5	88,9
	12000	26	6,9	95,8
	15000	1	,3	96,0
	20000	15	4,0	100,0
	Total	379	100,0	

Fuente. Elaboración propia.

En la Figura 23 se consideró el despliegue de las dos (2) variables rendimiento de pago y capital mediante un gráfico de dispersión, donde se evidenció la orientación de los resultados.

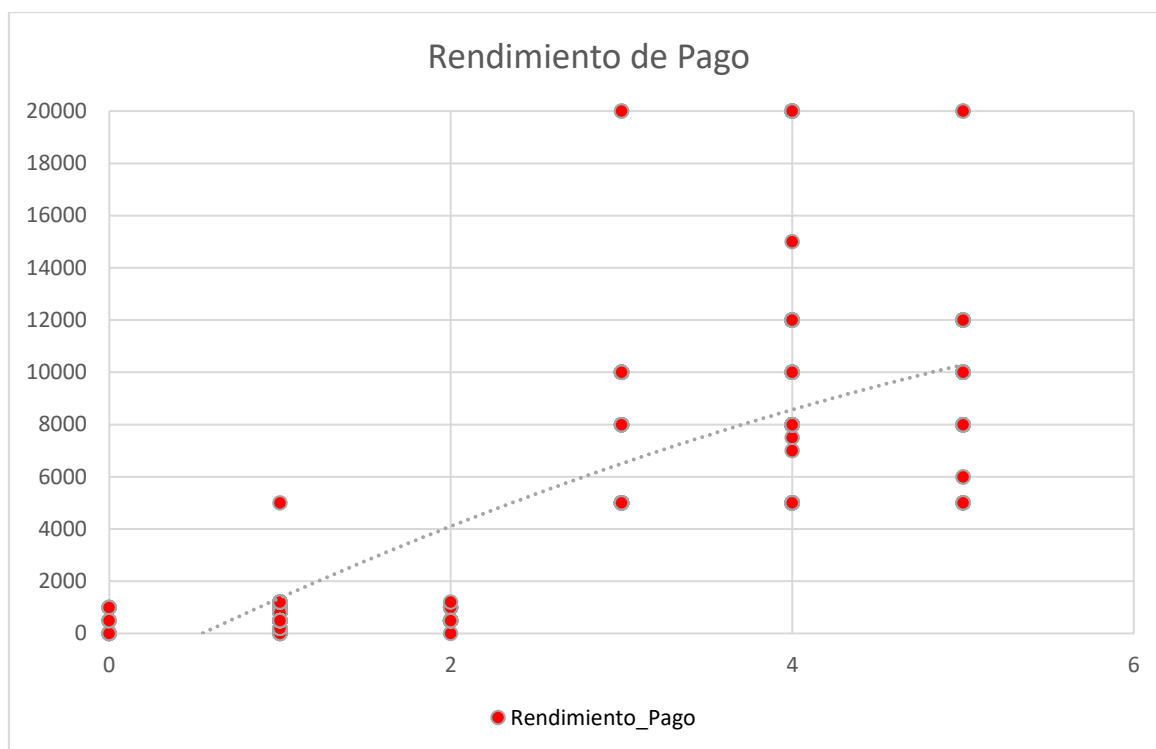


Figura 23. Resultados rendimiento de pago con capital

Fuente. Elaboración propia.

La rentabilidad es una variable muy importante para considerar por toda empresa porque permite identificar la eficiencia en la administración de los ingresos. Asimismo, se ha identificado que cuando una empresa tiene un mayor tiempo de actividad presenta una mayor rentabilidad. Motivo por el cual, es muy posible que se origine una quiebra desde los primeros meses de actividad.

En la Tabla 28 se precisaron resultados generados sobre rentabilidad mediante la estadística descriptiva, los cuales fueron:

Tabla 28. Tabla de frecuencias de la variable rentabilidad

Rentabilidad		F	P%	P% válido	P% acumulado
Válido	0	128	33,8	33,8	33,8
	1	6	1,6	1,6	35,4
	3	4	1,1	1,1	36,4
	5	241	63,6	63,6	100,0
	Total	379	100,0	100,0	

Fuente. Elaboración propia.

En la Tabla 29 se precisaron resultados generados sobre tiempo de actividad mediante la estadística descriptiva, los cuales fueron:

Tabla 29. Tabla de frecuencias de la variable tiempo de actividad

Tiempo de actividad		F	P%	P% válido	P% acumulado
Válido	0	6	1,6	1,6	1,6
	1	34	9,0	9,0	10,6
	2	57	15,0	15,0	25,6
	3	24	6,3	6,3	31,9
	4	2	,5	,5	32,5
	5	8	2,1	2,1	34,6
	6	4	1,1	1,1	35,6
	8	2	,5	,5	36,1
	9	1	,3	,3	36,4
	10	31	8,2	8,2	44,6
	12	70	18,5	18,5	63,1
	16	106	28,0	28,0	91,0
	18	34	9,0	9,0	100,0
	Total	379	100,0	100,0	

Fuente. Elaboración propia.

En la Figura 24 se consideró el despliegue de las dos (2) variables rentabilidad y tiempo de actividad mediante un gráfico de dispersión, donde se evidenció la orientación de los resultados.

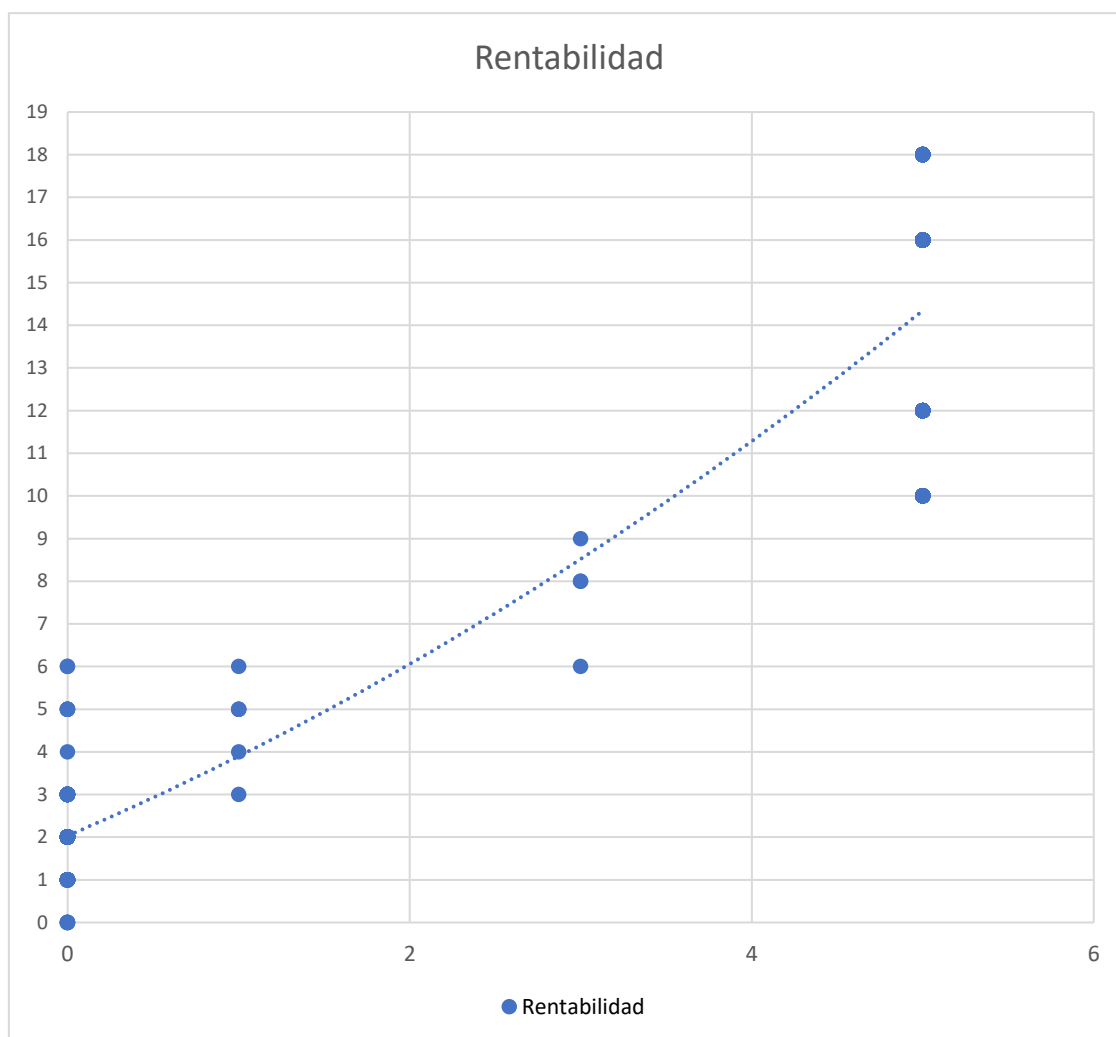


Figura 24. Resultados rentabilidad con tiempo de actividad (meses)
Fuente. Elaboración propia.

La satisfacción empresarial permite identificar aspectos determinantes no solo del clima laboral en la empresa sino más bien del entorno sobre el contexto de consolidación y salud monetaria; con ello se ha obtenido que el grado de satisfacción es mayor cuando la empresa se va consolidando en el tiempo. Caso contrario a las empresas que tienen menor tiempo de actividad donde se identificó que tienen una satisfacción menor y que pueden originar una quiebra.

En la Tabla 30 se precisaron resultados generados por satisfacción empresarial mediante la estadística descriptiva, los cuales fueron:

Tabla 30. Tabla de frecuencias de la variable satisfacción empresarial

Satisfacción Empresarial	F	P%	P% válido	P% acumulado
Válido	0	94	24,8	24,8
	1	40	10,6	35,4
	3	29	7,7	43,0
	4	163	43,0	86,0
	5	53	14,0	100,0
	Total	379	100,0	

Fuente. Elaboración propia.

En la Tabla 31 se precisaron resultados generados para tiempo de actividad mediante la estadística descriptiva.

Tabla 31. Tabla de frecuencias de la variable tiempo de actividad

Tiempo de actividad	F	P%	P% válido	P% acumulado
Válido	0	6	1,6	1,6
	1	34	9,0	10,6
	2	57	15,0	25,6
	3	24	6,3	31,9
	4	2	,5	32,5
	5	8	2,1	34,6
	6	4	1,1	35,6
	8	2	,5	36,1
	9	1	,3	36,4
	10	31	8,2	44,6
	12	70	18,5	63,1
	16	106	28,0	91,0
	18	34	9,0	100,0
	Total	379	100,0	

Fuente. Elaboración propia.

En la Figura 25 se consideró el despliegue de las variables satisfacción empresarial y tiempo de actividad mediante un gráfico de dispersión, donde se evidenció la orientación de los resultados.

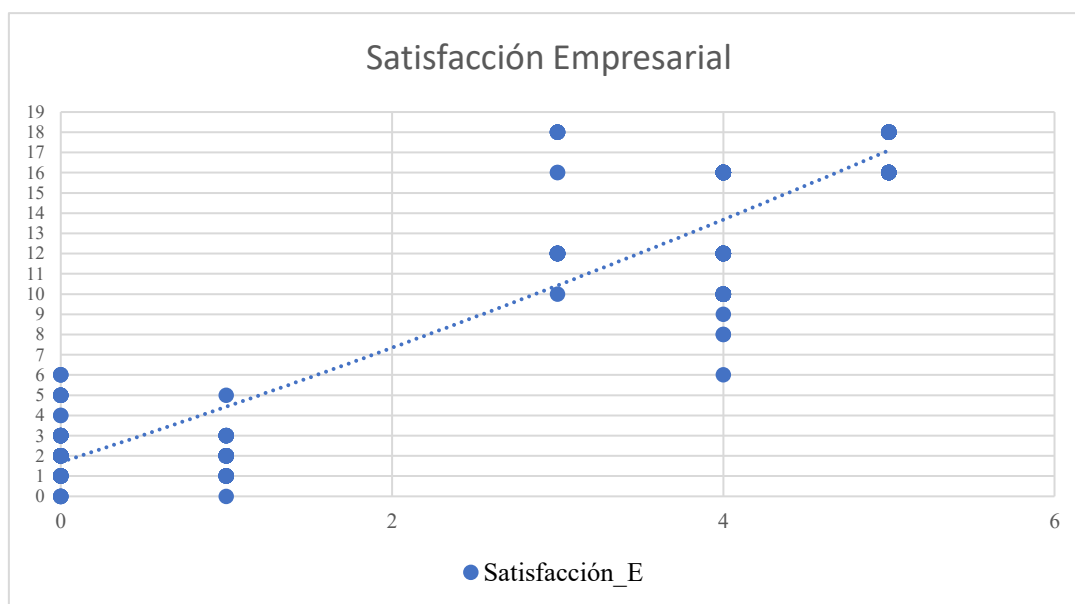


Figura 25. Resultados satisfacción empresarial con tiempo de actividad
Fuente. Elaboración propia.

La insolvencia y el tiempo de actividad son variables importantes dentro de todo modelo de quiebra empresarial. Se identificó que hay una mayor insolvencia entre las empresas que tienen menos tiempo de actividad, y la insolvencia menor se encuentra con las empresas que tienen mayor tiempo de actividad. En la Tabla 32 se precisaron resultados generados para insolvencia mediante la estadística descriptiva, los cuales fueron:

Tabla 32. Tabla de frecuencias insolvencia

Insolvencia		F	P%	P% válido	P% acumulado
Válido	0	159	42,0	42,0	42,0
	1	51	13,5	13,5	55,4
	2	35	9,2	9,2	64,6
	4	50	13,2	13,2	77,8
	5	84	22,2	22,2	100,0
	Total	379	100,0	100,0	

Fuente. Elaboración propia.

En la Tabla 33 se precisaron resultados obtenidos para la variable tiempo de actividad mediante la aplicación de la estadística descriptiva.

Tabla 33. Tabla de frecuencias de la variable tiempo de actividad

Tiempo de actividad	F	P%	P% válido	P% acumulado
0	6	1,6	1,6	1,6
1	34	9,0	9,0	10,6
2	57	15,0	15,0	25,6
3	24	6,3	6,3	31,9
4	2	,5	,5	32,5
5	8	2,1	2,1	34,6
6	4	1,1	1,1	35,6
8	2	,5	,5	36,1
9	1	,3	,3	36,4
10	31	8,2	8,2	44,6
12	70	18,5	18,5	63,1
16	106	28,0	28,0	91,0
18	34	9,0	9,0	100,0
Total	379	100,0	100,0	

Fuente. Elaboración propia.

En la Figura 26 se consideraron las variables insolvencia y tiempo de actividad con un gráfico de dispersión, donde se evidenció la orientación de los resultados.

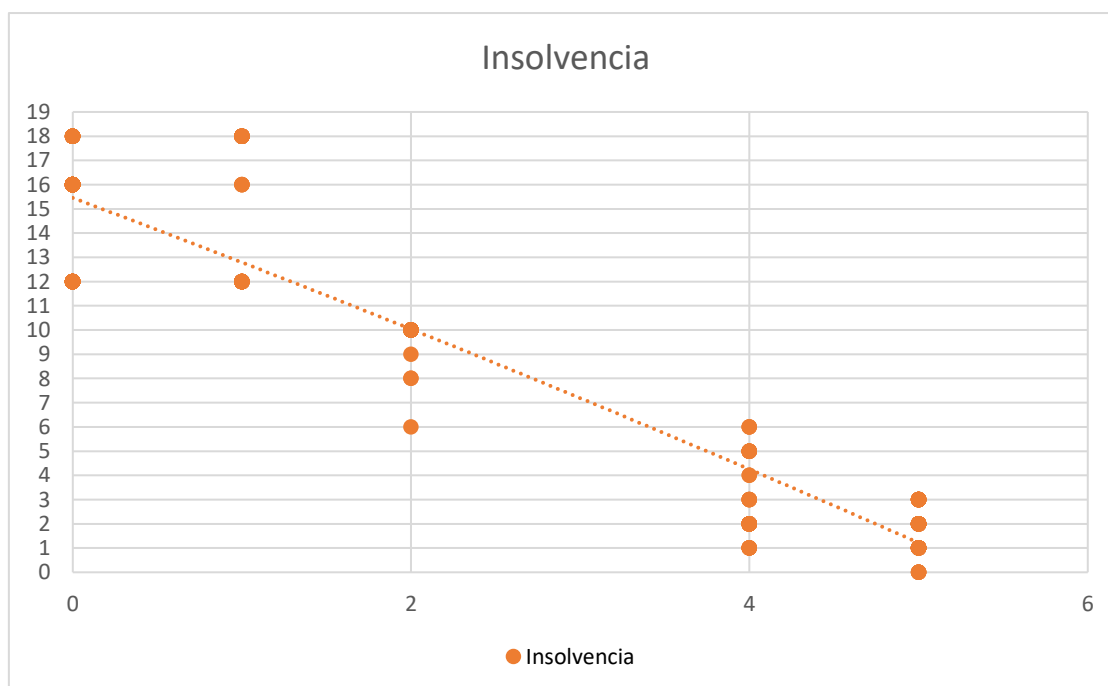


Figura 26. Resultados insolvencia y tiempo de actividad (meses)

Fuente. Elaboración propia.

4.1.4. Análisis de resultados de las métricas de performance del modelo optimizado

Se procedió al despliegue del conjunto de datos, donde se obtuvo una precisión de 97%, exactitud de 97%, sensibilidad de 99% y f1-score de 98%. En la Tabla 34 se precisaron los resultados de la matriz de confusión del modelo optimizado.

Tabla 34. Matriz de confusión del modelo optimizado

Quiebra de empresas	Predicción			Total
	Positivo	Negativo	Porcentaje Correcto	
Positivo	122	8	32%	379
Negativo	4	245	65%	
	33%	67%	97%	

Fuente. Elaboración propia.

Los resultados que se obtuvieron en la matriz de confusión determinaron que existe un porcentaje de 97% que representan a 367 empresas que obtuvieron una predicción correcta (verdaderos positivos y negativos). Sin embargo, se ha identificado que el 3% (falsos positivos y negativos) corresponde a 12 empresas que obtuvieron un pronóstico con error. En ese sentido, es importante que la minimización de los errores del modelo se realice de manera eficiente, donde los falsos positivos y negativos representan eventos críticos que generan incertidumbre y es necesario que puedan reducirse. Asimismo, se identificaron 8 empresas como falsos negativos, los cuales son fallos peligrosos del modelo porque pueden generar problemas de endeudamiento o quiebra temprana al considerar empresas que podrían obtener préstamos o panoramas positivos cuando su realidad es diferente. Es importante identificar a las empresas para el tratamiento adecuado mediante un seguimiento y control que brinde un panorama claro. Representa un impacto importante tanto para las empresas como para el modelo propuesto.

Los resultados generados por las métricas del modelo optimizado se precisaron en la Tabla 35, los cuales fueron necesarios para la evaluación del modelo.

Tabla 35. Resultados del modelo optimizado

	precision	sensibilidad	f1-score	Total
	0.97	0.99	0.98	379
0	0.32	0.41	0.46	
1	0.65	0.58	0.52	
exactitud		0.97	379	
med-prom	0.74	0.69	0.71	379
pond-prom	0.85	0.87	0.86	379

Fuente. Elaboración propia.

El código ejecutado para el modelo optimizado se consideró con la sentencia `accuracy_score` (Python), donde se determinó el 97.37% de exactitud.

```
ANN_accuracy= accuracy_score(ytest, ypred)
```

```
print("El Accuracy del ANN es", ANN_accuracy*100,"%")
```

Asimismo, en los resultados obtenidos se identificaron el número de empresas que experimentarían una quiebra.

Total = 379

Número de empresas que tienen riesgo de quiebra = 134

Porcentaje de empresas que tienen riesgo de quiebra = 35.35620052770449%

Número de empresas que no tienen riesgo de quiebra = 245

Porcentaje de empresas que no tienen riesgo de quiebra = 64.64379947229551%

La prueba de la curva ROC precisa la clasificación binaria de las empresas para los grupos positivos y negativos. En ese sentido, en la Figura 27 se aprecian los resultados obtenidos que demuestran el rendimiento del modelo tiene una mayor área contenida con predominio hacia la izquierda y donde se precisa el valor del clasificador AUC de 0.97 (cercano a 1).

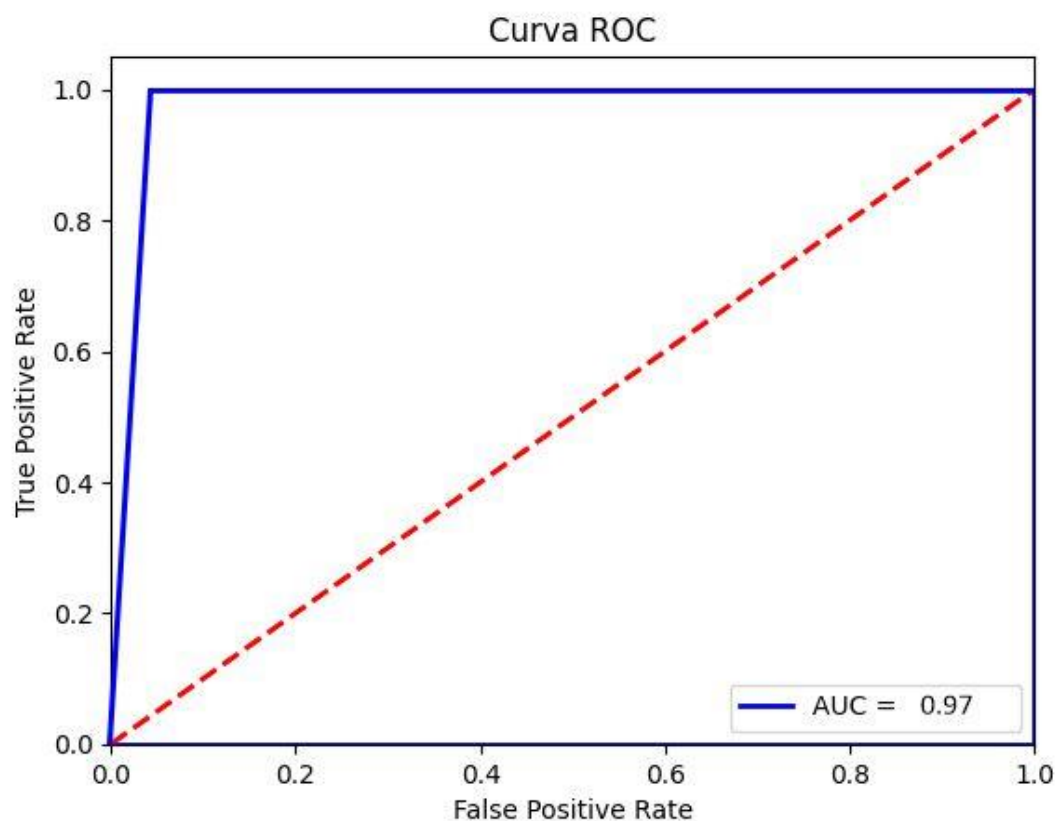


Figura 27. Resultados prueba curva ROC

Fuente. Elaboración propia.

Se realizó la prueba de validación cruzada (KFold) para evaluar el rendimiento del modelo optimizado y generar una garantía sobre los resultados obtenidos, según las pruebas de entrenamiento para la matriz del conjunto de datos. En ese sentido, se realizó la validación cruzada con $k=5$; es decir, se dividió al conjunto de datos en cinco (5) partes y donde cada parte fue considerada como el conjunto de prueba y de entrenamiento según las iteraciones realizadas.

En ese contexto, para R^2 se obtuvo: Parte 1 (0.94), Parte 2 (0.92), Parte 3 (1), Parte 4 (0.965) y Parte 5 (0.97), donde el conjunto de datos para cada parte fue: Parte 1, 2, 3 y 4 fue de 76, y la Parte 5 fue de 75. En la Figura 28, se aprecia los resultados obtenidos:

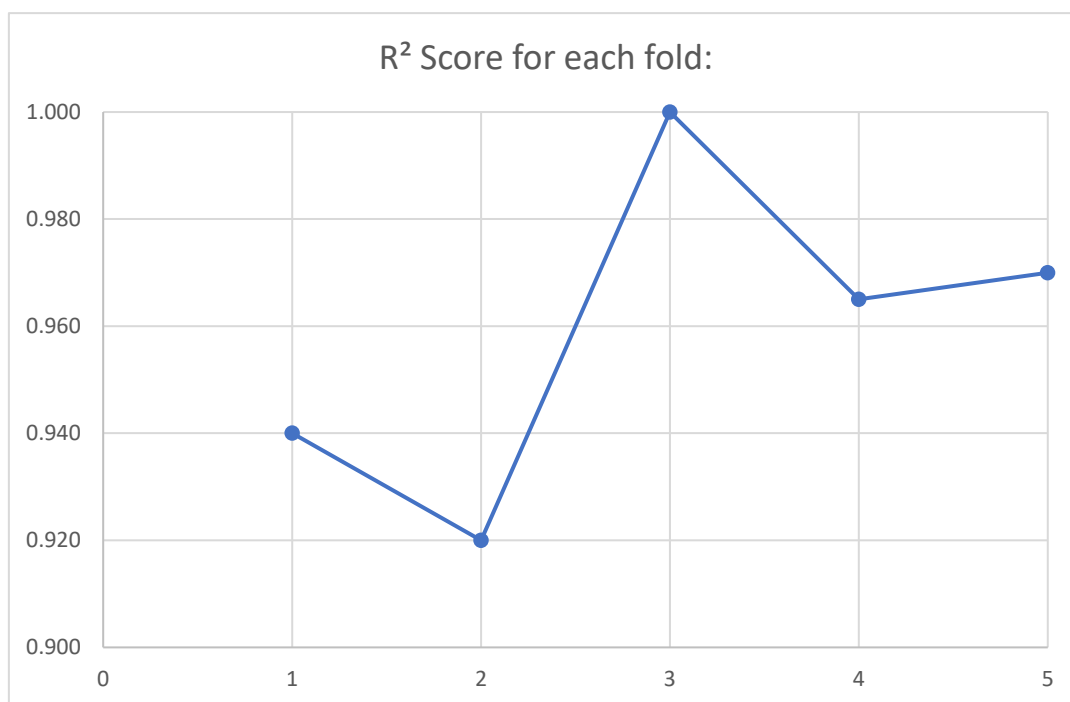


Figura 28. Validación cruzada KFold

Fuente. Elaboración propia.

Se identificó que el promedio R^2 fue de: 0.959 (95.9%). El cual significa que el rendimiento del modelo es muy confiable y ofrece una evaluación más robusta en el desempeño.

Se realizó el ajuste de hiperparámetros para mejorar el rendimiento del modelo optimizado mediante la optimización bayesiana, en la Tabla 36 se visualiza los resultados siguientes:

Tabla 36. Ajuste hiperparámetros

Hiperparámetros	% Optimización
Meses_A	80%
Capital	87%
Satisfacción_E	82%
Acceso_Credito	34%
Letras_por_pagar	42%
Rendimiento_Pago	85%
Rentabilidad	85%
Insolvencia	85%

Fuente. Elaboración propia.

Los resultados de las métricas generadas por la optimización bayesiana fueron precisados en la Tabla 37. Los cuales fueron necesarios para identificar la pertinencia de mejoramiento del modelo.

Tabla 37. Resultados de la optimización bayesiana

	precision	sensibilidad	f1-score	Total
	0.95	0.96	0.96	379
0	0.29	0.40	0.45	
1	0.66	0.56	0.51	
exactitud		0.96	379	
med-prom	0.91	0.90	0.91	379
pond-prom	0.89	0.92	0.93	379

Fuente. Elaboración propia.

Se compararon los resultados obtenidos por el modelo optimizado y la optimización bayesiana, los cuales se precisaron en la Tabla 38. Se obtuvo que el modelo optimizado supera a la optimización bayesiana en los resultados obtenidos por las métricas aplicadas.

Tabla 38. Comparación de optimización

Métrica	MO	OB
Precisión	97%	95%
Exactitud	97%	96%
Sensibilidad	99%	96%
F1-Score	98%	96%

Fuente. Elaboración propia.

Los resultados obtenidos luego de la aplicación de la técnica de selección de características (RFE) fueron precisadas en la Figura 29, donde se identificó que las variables Insolvencia (90%), Rendimiento_Pago (85%), Capital (89%), Rentabilidad (86%), Satisfacción (89%) y Meses_A (90%) son aquellas tienen mayor impacto para el modelo según el porcentaje obtenido.

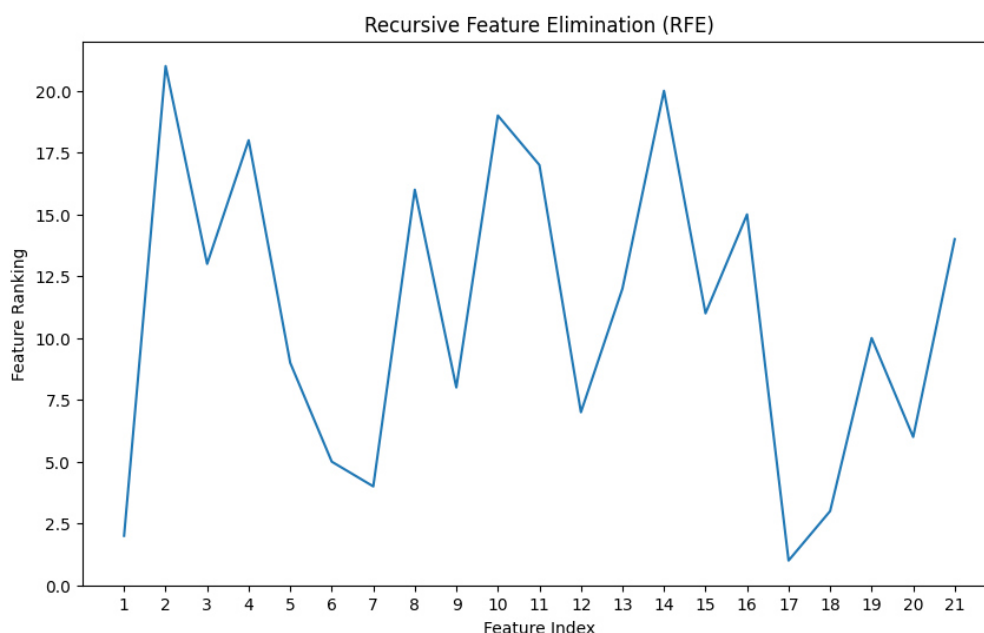


Figura 29. RFE

Fuente. Elaboración propia.

4.2. Pruebas de hipótesis

La investigación consideró la comprobación de las hipótesis (hipótesis general y específicas) mediante métodos estadísticos determinados según la estructura y las actividades de cada objetivo. Los cuales se realizaron de la siguiente manera:

Hipótesis 1 (H1): El preprocesamiento de la data colectada de empresas optimiza significativamente la predicción del modelo de aprendizaje automático.

Hipótesis Nula (H0): El preprocesamiento de la data colectada de empresas no optimiza significativamente la predicción del modelo de aprendizaje automático.

La prueba de normalidad de Kolmogorov-Smirnov fue realizada con una muestra de 379 empresas para identificar si los datos se distribuyen con normalidad. En ese sentido, los datos correspondieron a la matriz del conjunto de datos antes y después del preprocesamiento para determinar posibles diferencias entre ellas. En la siguiente Tabla 39 se puede apreciar que respecto a la prueba de normalidad no se presentaron diferencias significativas entre ambos cálculos.

Tabla 39. Prueba de normalidad

	Kolmogorov-Smirnov ^a		
	Estadístico	gl	Sig.
Quiebra_LB	.515	379	.000
Quiebra_PP	.497	379	.000

Fuente. Elaboración propia.

La prueba de normalidad realizada generó resultados de Significancia menores de 0.05, motivo por la cual no se presentó una distribución normal (no paramétrica). En ese sentido, para comprobar la hipótesis la medida estadística que se empleó fue la de Wilcoxon. Los datos para la prueba se obtuvieron de 2 grupos relacionados (antes y después del preprocesamiento) de la matriz del conjunto de datos, los cuales consideran el nivel ordinal de las variables. En la Tabla 40 se presentaron la asignación con rangos (empate, positivo y negativo) y la suma de rangos (positivo y negativo). Asimismo, se evidenció que existen más rangos positivos que negativos para el conjunto de datos luego del preprocesamiento.

Tabla 40. Prueba con rangos Wilcoxon

		N	Rango promedio	Suma de rangos
Quiebra_PP	Rangos negativos	28 ^a	36,50	1022,00
Quiebra_LB	Rangos positivos	44 ^b	36,50	1606,00
	Empates	307 ^c		
	Total	379		

a. Quiebra_PP < Quiebra_LB

b. Quiebra_PP > Quiebra_LB

c. Quiebra_PP = Quiebra_LB

Fuente. Elaboración propia.

Se consideró la prueba Wilcoxon en la Tabla 41, donde se obtuvo como resultado de significancia el 0.045. En ese sentido, la hipótesis nula (H0) fue rechazada y la hipótesis alterna (H1) fue aprobada. Además, el valor de significancia fue menor a 0.05. Por lo tanto, se presentaron diferencias significativas entre la matriz del conjunto de datos antes y después del preprocesamiento.

Tabla 41. Prueba no paramétrica Wilcoxon

	Quiebra_PP Quiebra_LB
Z	-1,886 ^b
Sig.	.045

Fuente. Elaboración propia.

Finalmente, se brindó validez y comprobación a la H1: El preprocesamiento de la data colectada de empresas optimiza significativamente la predicción del modelo de aprendizaje automático.

Hipótesis 2 (H2): La evaluación del performance de los algoritmos de aprendizaje automático optimiza significativamente la predicción del modelo de la quiebra de empresas.

Hipótesis Nula (H0): La evaluación del performance de los algoritmos de aprendizaje automático no optimiza significativamente la predicción del modelo de la quiebra de empresas.

Para validar las hipótesis 2, se realizó el tratamiento con Kappa de Cohen, requerido para validar los resultados de concordancia entre dos clasificadores y obtener el grado de significancia. Se precisaron resultados en la Tabla 42, donde el mayor valor del índice de precisión fue del algoritmo red neuronal con 90%.

Tabla 42. Índice de precisión con los algoritmos evaluados

Nro	Algoritmo	Precisión
1	XGBoost	88%
2	SVM	79%
3	LSTM	75%
4	Regresión logística	85%
5	Bosques aleatorios	83%
6	Red neuronal	90%

Fuente. Elaboración propia.

En la Tabla 43 se precisó que el mayor valor del índice de exactitud fue del algoritmo red neuronal con 87%.

Tabla 43. Índice de exactitud con los algoritmos evaluados

Nro	Algoritmo	Exactitud
1	XGBoost	86%
2	SVM	78%
3	LSTM	74%
4	Regresión logística	83%
5	Bosques aleatorios	83%
6	Red neuronal	87%

Fuente. Elaboración propia.

Se precisaron resultados en la Tabla 44, donde el mayor valor del índice de sensibilidad fue del algoritmo red neuronal con 97%.

Tabla 44. Índice de sensibilidad con los algoritmos evaluados

Nro	Algoritmo	Sensibilidad
1	XGBoost	92%
2	SVM	85%
3	LSTM	82%
4	Regresión logística	95%
5	Bosques aleatorios	96%
6	Red neuronal	97%

Fuente. Elaboración propia.

En la Tabla 45 se precisó que el mayor valor del índice de f1-score fue del algoritmo red neuronal con 92%.

Tabla 45. Índice de f1-score con los algoritmos evaluados

Nro	Algoritmo	F1-Score
1	XGBoost	90%
2	SVM	84%

3	LSTM	84%
4	Regresión logística	90%
5	Bosques aleatorios	90%
6	Red neuronal	92%

Fuente. Elaboración propia.

Se aprecia en la Figura 30 que la red neuronal obtuvo los mejores resultados respecto todos los otros algoritmos con precisión de 90%, exactitud de 87%, sensibilidad de 97% y f1-score de 92%.

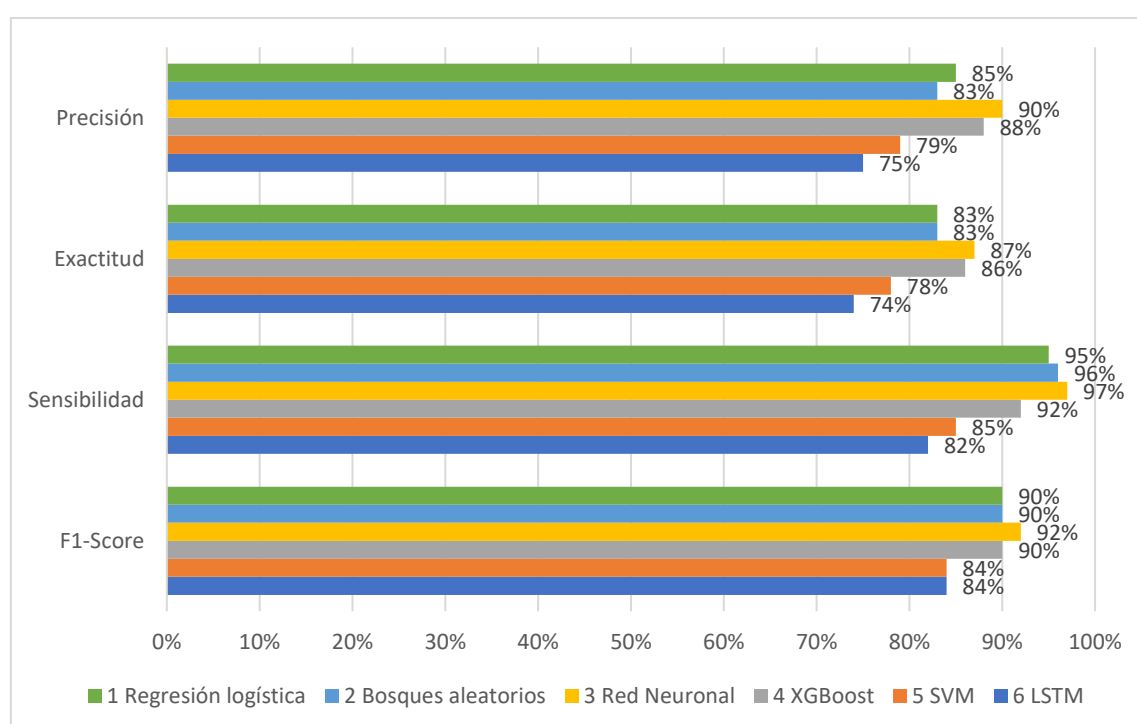


Figura 30. Resultados de las métricas

Fuente. Elaboración propia.

Para conocer y validar la concordancia que permita medir la significancia entre dos clasificadores se consideró el coeficiente Kappa de Cohen. Los autores Mobbs et al. (2024) precisaron que el coeficiente muestra los niveles de concordancia entre dos observadores y que los resultados pueden generar valores del intervalo de -1 y +1. Asimismo, el valor más cercano al +1 permitirá un grado mayor de concordancia. En la Tabla 46 se puede apreciar los valores del rango considerados

para la evaluación del coeficiente Kappa de Cohen, el cual tiene diversos niveles según correspondan los resultados.

Tabla 46. Valoración de concordancia

Coeficiente Kappa	Fuerza de la concordancia
< 0.01	Pobre
0.01-0.20	Leve
0.21-0.40	Aceptable
0.41-0.60	Moderada
0.61-0.80	Considerable
0.81-1.00	Casi perfecta

Fuente: Mobbs et al. (2024)

Los resultados generados por la matriz de confusión de cada algoritmo evaluado fueron los siguientes. En la Tabla 47 se aprecia el valor obtenido con la medición de Kappa donde se obtuvo 0.80; por la cual, se considera que se presenta concordancia considerable entre los clasificadores del algoritmo. Asimismo, se obtuvo significancia menor a 0.05; se aprobó la concordancia referida al algoritmo XGBoost.

Tabla 47. Medidas: Kappa de Cohen – XGBoost

		Medidas-simétricas			
		Val	Error estándar asintótico	T aprox	Sig
Med	Kappa	0.80	0.78	0.84	.000
Casos válidos		379			

Fuente. Elaboración propia.

En la Tabla 48 se aprecia el valor obtenido con la medición de Kappa donde se obtuvo 0.69; por la cual, se considera que se presenta concordancia considerable entre los clasificadores del algoritmo. Asimismo, se obtuvo significancia menor a 0.05; se aprobó la concordancia referida al algoritmo SVM.

Tabla 48. Medidas: Kappa de Cohen – SVM

		Medidas-simétricas			
		Val	Error estándar asintótico	T aprox	Sig
Med	Kappa	0.69	0.60	0.70	.000
Casos válidos		379			

Fuente. Elaboración propia.

En la Tabla 49 se aprecia el valor obtenido con la medición de Kappa donde se obtuvo 0.79; por la cual, se considera que se presenta concordancia considerable entre los clasificadores del algoritmo. Asimismo, se obtuvo significancia menor a 0.05; se aprobó la concordancia referida al algoritmo LSTM.

Tabla 49. Medidas: Kappa de Cohen – LSTM

		Medidas-simétricas			
		Val	Error estándar asintótico	T aprox	Sig
Med	Kappa	0.79	0.72	0.83	.000
Casos válidos		379			

Fuente. Elaboración propia.

En la Tabla 50 se aprecia el valor obtenido con la medición de Kappa donde se obtuvo 0.70; por la cual, se considera que se presenta concordancia considerable entre los clasificadores del algoritmo. Asimismo, se obtuvo significancia menor a 0.05; se aprobó la concordancia referida al algoritmo bosques aleatorios.

Tabla 50. Medidas: Kappa de Cohen – Bosques aleatorios

		Medidas-simétricas			
		Val	Error estándar asintótico	T aprox	Sig
Med	Kappa	0.70	0.65	0.78	.000
Casos válidos		379			

Fuente. Elaboración propia.

En la Tabla 51 se aprecia el valor obtenido con la medición de Kappa donde se obtuvo 0.86; por la cual, se considera que se presenta concordancia casi perfecta entre los clasificadores del algoritmo. Asimismo, se obtuvo significancia menor a 0.05; se aprobó la concordancia referida al algoritmo regresión logística.

Tabla 51. Medidas: Kappa de Cohen – Regresión logística

		Medidas-simétricas			
		Val	Error estándar asintótico	T aprox	Sig
Med	Kappa	0.86	0.91	0.98	.000
Casos válidos		379			

Fuente. Elaboración propia.

En la Tabla 52 se aprecia el valor obtenido con la medición de Kappa donde se obtuvo 0.88; por la cual, se considera que se presenta concordancia casi perfecta entre los clasificadores del algoritmo. Asimismo, se obtuvo significancia menor a 0.05, se aprobó la concordancia referida al algoritmo red neuronal.

Tabla 52. Medidas: Kappa de Cohen – Red neuronal

		Medidas-simétricas			
		Val	Error estándar asintótico	T aprox	Sig
Med	Kappa	0.88	0.85	0.91	.000
Casos válidos		379			

Fuente. Elaboración propia.

Los algoritmos que obtuvieron un mejor nivel de concordancia fueron: Regresión logística y red neuronal con >0.81 según las medidas de acuerdo con la tabla de medición de Kappa de Cohen, siendo el algoritmo red neuronal quien obtuvo un valor mayor 0.88 de concordancia. Además, obtuvo valor menor de 0.05 en significancia. En ese sentido, se rechazó la hipótesis nula (H_0) para aprobar la hipótesis alterna (H_2). Finalmente, se brindó validez y comprobación a la H_2 : La

evaluación del performance de los algoritmos de aprendizaje automático optimiza significativamente la predicción del modelo de la quiebra de empresas.

Hipótesis 3 (H3): La correlación de los factores relevantes de las empresas optimiza significativamente el diseño del modelo de aprendizaje automático de la predicción de quiebra.

Hipótesis Nula (H0): La correlación de los factores relevantes de las empresas no optimiza significativamente el diseño del modelo de aprendizaje automático de la predicción de quiebra.

Para la comprobación de la hipótesis se consideraron la identificación de variables relevantes para la quiebra de empresas y se generaron correlaciones exclusivas de variables más idóneas. Posterior a ello, se realizó un proceso de implementación y entrenamiento en el modelo optimizado. La prueba de normalidad de Kolmogorov-Smirnov fue realizada con una muestra de 379 empresas para identificar si los datos se distribuyen con normalidad. En ese sentido, los datos correspondieron a la matriz del conjunto de datos después del preprocesamiento para determinar posibles diferencias entre las variables.

En la siguiente Tabla 53 se puede apreciar el cálculo respecto a la prueba de normalidad entre las variables rendimiento y capital. La prueba de normalidad consideró un valor menor a 0.05 como significancia; en ese sentido, se determinó que no tiene una distribución normal (no paramétrica). Asimismo, se realizó la correlación de variables mediante la prueba de correlación Rho Spearman.

Tabla 53. Prueba de normalidad

	Kolmogorov-Smirnov ^a		
	Estadístico	gl	Sig.
Rendimiento_Pago	.212	379	.000
Capital	.181	379	.000

Fuente. Elaboración propia.

En la Tabla 54 se precisaron resultados de la prueba Rho Spearman donde se generó el coeficiente de correlación. En ese marco, se identificó el nivel de correlación entre las variables rendimiento de pago y el capital. Se obtuvo que $p = .000 < \alpha = .01$; motivo por el cual, se presentó una relación altamente significativa entre el rendimiento de pago y el capital. Por lo tanto, según la empresa tiene mayor capital entonces va a obtener un mayor rendimiento de pago para cumplir sus obligaciones a mediano y largo plazo. El valor $Rho = .492$ precisa correlación positiva moderada.

Tabla 54. Correlación entre rendimiento y capital

Correlaciones		
	Rendimiento	Capital
Coefcorrel	1.000	.492**
Signb	-	.000
N.ro	379	379
Coefcorrel	.492**	1.000
Signb	.000	-
N.ro	379	379
**. Correl signif.		

Fuente. Elaboración propia.

En la siguiente Tabla 55 se puede apreciar el cálculo respecto a la prueba de normalidad para las variables rentabilidad con tiempo de actividad. La prueba de normalidad consideró un valor menor a 0.05 como significancia; en ese sentido, se determinó que no tiene una distribución normal (no paramétrica). Asimismo, se realizó la correlación de variables mediante la prueba de correlación Rho Spearman.

Tabla 55. Prueba de normalidad

	Kolmogorov-Smirnov ^a		
	Estadístico	gl	Sig.
Rentabilidad	.205	379	.000
Tiempo de actividad	.170	379	.000

Fuente. Elaboración propia.

En la Tabla 56 se precisaron resultados de la prueba Rho Spearman donde se generó el coeficiente de correlación. En ese marco, se identificó el nivel de correlación entre las variables rentabilidad y tiempo de actividad. Se obtuvo que $p = .000 < \alpha = .01$; motivo por el cual, se presentó una relación altamente significativa entre la rentabilidad y el tiempo de actividad. Por lo tanto, según se tiene mayor tiempo de actividad se generará mayor rentabilidad. El valor de Rho = .392 precisa una correlación positiva baja.

Tabla 56. Correlación entre rentabilidad con tiempo de actividad

Correlaciones		
	Rentabilidad	Tiempo de actividad
Coefcorrel	1.000	.392**
Signb	-	.000
N.ro	379	379
Coefcorrel	.392**	1.000
Signb	.000	-
N.ro	379	379
**. Correl signif.		

Fuente. Elaboración propia.

En la Tabla 57 se puede apreciar el cálculo respecto a la prueba de normalidad para las variables satisfacción empresarial y tiempo de actividad. La prueba de normalidad consideró un valor menor a 0.05 como significancia; en ese sentido, se determinó que no tiene una distribución normal. Asimismo, se realizó la correlación de variables mediante la prueba de correlación Rho Spearman.

Tabla 57. Prueba de normalidad

	Kolmogorov-Smirnov ^a		
	Estadístico	gl	Sig.
Satisfacción_E	.211	379	.000
Tiempo de actividad	.192	379	.000

Fuente. Elaboración propia.

En la Tabla 58 se precisaron resultados de la prueba Rho Spearman donde se generó el coeficiente de correlación. En ese marco, se identificó el nivel de correlación entre las variables satisfacción y el tiempo de actividad. Se obtuvo $p = .006 < \alpha = .01$. Motivo por el cual, se presentó una relación altamente significativa entre satisfacción y tiempo de actividad. Por lo tanto, según la empresa tiene más tiempo de actividad presentará mayor satisfacción empresarial. El valor de $Rho = .216$ precisa una correlación positiva baja. Se realizó la prueba de normalidad para las variables insolvencia con tiempo de actividad

Tabla 58. Correlación entre satisfacción y tiempo de actividad

Correlaciones		
	Satisfacción empresarial	Tiempo de actividad
Coefcorrel	1.000	.216**
Signb	-	.006
N.ro	379	379
Coefcorrel	.216**	1.000
Signb	.006	-
N.ro	379	379
**. Correl signif.		

Fuente. Elaboración propia.

En la Tabla 59 se puede apreciar el cálculo respecto a la prueba de normalidad para las variables insolvencia y tiempo de actividad. La prueba de normalidad consideró un valor menor a 0.05 como significancia; en ese sentido, se determinó que no tiene una distribución normal (no paramétrica). Asimismo, se realizó la correlación de variables mediante la prueba de correlación Rho Spearman.

Tabla 59. Prueba de normalidad

	Kolmogorov-Smirnov ^a		
	Estadístico	gl	Sig.
Insolvencia	.208	379	.000
Tiempo de actividad	.188	379	.000

Fuente. Elaboración propia.

En la Tabla 60 se precisaron resultados de la prueba Rho Spearman donde se generó el coeficiente de correlación. En ese marco, se identificó el nivel de correlación entre las variables insolvencia y el tiempo de actividad. Se obtuvo que $p = .000 < \alpha = .01$; motivo por el cual, se presentó una relación altamente significativa entre la insolvencia y tiempo de actividad. Por lo tanto, se identifica relación entre las empresas que tienen menos tiempo de actividad con mayor insolvencia. El valor Rho = .210 precisa una correlación positiva baja.

Tabla 60. Correlación entre insolvencia y tiempo de actividad

Correlaciones		
	Insolvencia	Tiempo de actividad
Coefcorrel	1.000	.210**
Signb	-	.000
N.ro	379	379
Coefcorrel	.318**	1.000
Signb	.000	-
N.ro	379	379
**. Correl signif.		

Fuente. Elaboración propia.

Se identificó que las empresas que consolidan sus actividades en periodos de tiempo mayores son aquellas que tienen recursos para sobrellevar actividades negativas como momentos de incertidumbre por temas económicos o problemas externos. Asimismo, la insolvencia es una característica importante dentro de una empresa porque permite motivar inversiones futuras o detener actividades, por ello es necesario conocer el índice de insolvencia para comprometer actividades futuras. Se obtuvo que las correlaciones determinaron resultados de significancia menores al 0.05. Motivo por el cual, se rechazó la hipótesis nula (H_0) y se aprobó la hipótesis alterna (H_3).

Finalmente, se brindó validez y comprobación a la H_3 : La correlación de los factores relevantes de las empresas optimiza significativamente el diseño del modelo de aprendizaje automático de la predicción de quiebra.

Hipótesis 4 (H4): La evaluación con métricas de performance del modelo de aprendizaje automático valida significativamente la predicción de la quiebra de empresas.

Hipótesis Nula (H0): La evaluación con métricas de performance del modelo de aprendizaje automático no valida significativamente la predicción de la quiebra de empresas.

Para validar las hipótesis se utilizó el coeficiente de Kappa de Cohen para identificar cómo se desempeñaron los algoritmos, validar la concordancia y el grado de significancia entre dos clasificadores respectivamente. En la Tabla 61 se puede apreciar la evaluación realizada y los resultados obtenidos para cada algoritmo.

Tabla 61. Evaluación con métricas de performance

Nro	Algoritmo	Precisión	Exactitud	Sensibilidad	F1-Score
1	XGBoost	82%	82%	85%	86%
2	SVM	79%	78%	85%	84%
3	LSTM	75%	74%	82%	84%
4	Regresión logística	85%	83%	95%	90%
5	Bosques aleatorios	83%	83%	96%	90%
6	Red neuronal	90%	87%	97%	92%
7	Red neuronal (MO2A)	97%	97%	99%	98%

Fuente. Elaboración propia.

En la Figura 31 se identificó que el algoritmo red neuronal (MO2A) obtuvo los mejores resultados respecto todos los otros algoritmos con precisión de 97%, exactitud de 97%, sensibilidad de 99% y f1-score de 98%.

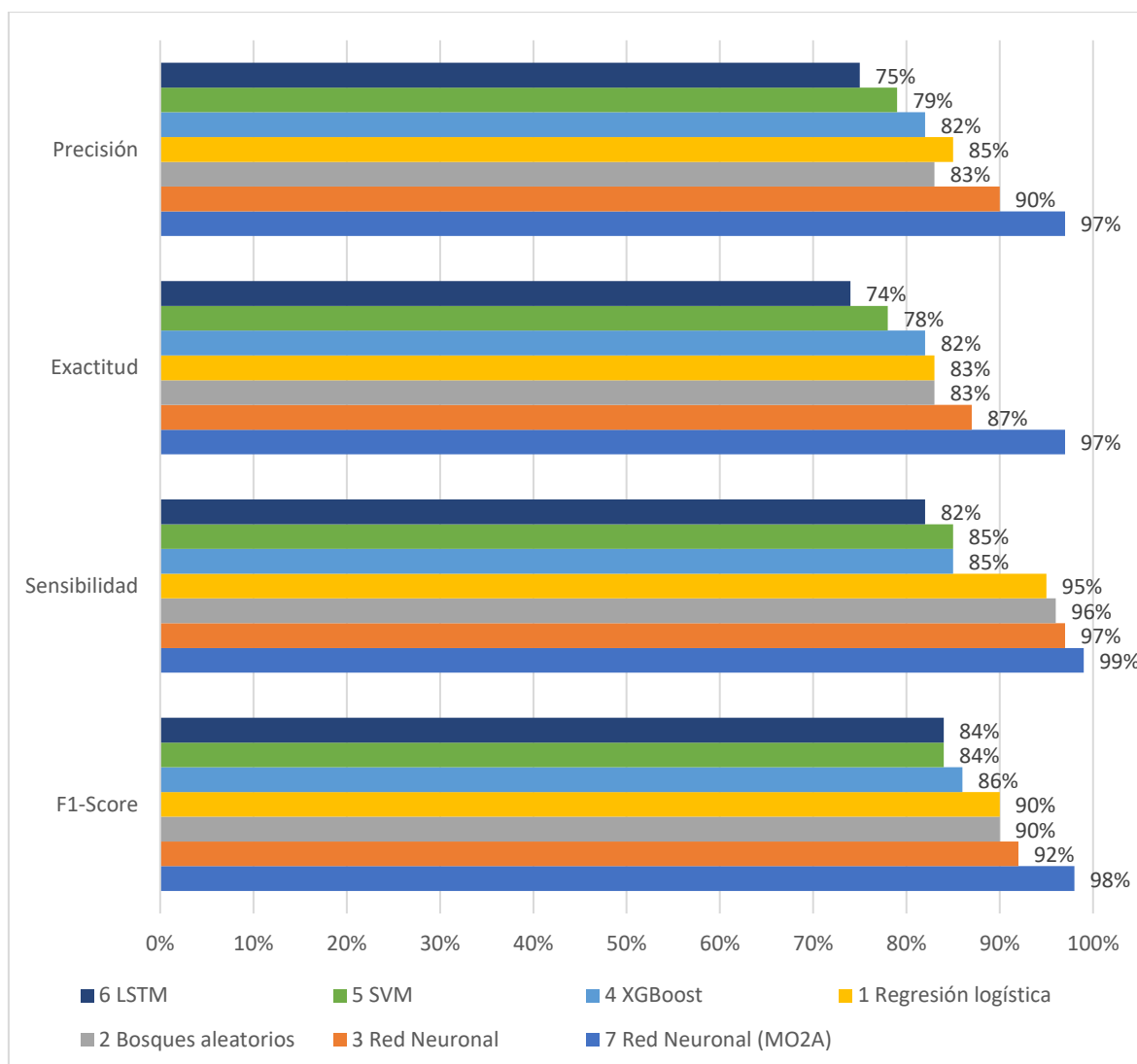


Figura 31. Resultados de las métricas

Fuente. Elaboración propia.

Asimismo, para conocer y validar la concordancia que permita medir la significancia entre dos clasificadores se consideró el coeficiente Kappa de Cohen. Los autores Mobbs et al. (2024) precisaron que el coeficiente muestra los niveles de concordancia entre dos observadores y que los resultados pueden generar valores

del intervalo de -1 y +1. Asimismo, el valor más cercano al +1 permitirá un grado mayor de concordancia. Los valores del rango considerados para la evaluación del coeficiente Kappa de Cohen fueron precisados en la valoración de la Tabla 46. En la Tabla 62 se puede apreciar los resultados generados por la matriz de confusión del algoritmo evaluado.

Tabla 62. Medidas: Kappa de Cohen – Red neuronal (MO2A)

		Medidas-simétricas			
		Val	Error estándar asintótico	T aprox	Sig
Med	Kappa	0.91	0.90	0.99	.000
Casos válidos		379			

Fuente. Elaboración propia.

El valor obtenido luego de la medición de Kappa fue 0.91; por la cual, se considera que se presenta concordancia casi perfecta entre los clasificadores del algoritmo. Asimismo, el valor fue menor a 0.05 respecto a la significancia; en ese sentido, se aprobó la hipótesis referida al algoritmo de red neuronal (MO2A). Se rechazó la hipótesis nula (H0) y se aprobó la hipótesis alterna (H4).

Finalmente, se brindó validez y comprobación a la H4: La evaluación con métricas de performance del modelo de aprendizaje automático valida significativamente la predicción de la quiebra de empresas.

Hipótesis General (HG): El desarrollo de un modelo de aprendizaje automático optimiza la predicción de la quiebra de empresas.

Hipótesis Nula (H0): El desarrollo de un modelo de aprendizaje automático no optimiza la predicción de la quiebra de empresas.

Para comprobar la hipótesis general se consideró la comprobación de las hipótesis específicas H1, H2, H3 y H4 según los métodos estadísticos realizados y los resultados obtenidos de significancia. Al respecto los autores Hernández-Sampieri

et al. (2018) consideraron que la hipótesis general son explicaciones afirmativas potencialmente válidas que se formulan mediante proposiciones. Asimismo, para validar las hipótesis es necesario generar una secuencia y relación entre las variables mediante métodos estadísticos especializados para las hipótesis específicas. De esta manera, se genera un rechazo sobre las hipótesis nulas y la aprobación de las alternas.

En la Tabla 63 se precisaron los resultados obtenidos por las cuatro (04) hipótesis específicas, los cuales fueron favorables y aprobadas; en su sentido, se rechazó la hipótesis nula (H_0) y se aprobó la hipótesis general (HG). Finalmente, se brindó validez y comprobación a la hipótesis general: El desarrollo de un modelo de aprendizaje automático optimiza la predicción de la quiebra de empresas.

Tabla 63. Comprobación de las hipótesis específicas

Hipótesis	Método Estadístico	Significancia	Validez	Situación
H1	Wilcoxon	.000	SI	Comprobado
H2	Kappa Cohen	.000	SI	Comprobado
H3	Rho-Spearman	.000	SI	Comprobado
H4	Kappa Cohen	.000	SI	Comprobado

Fuente. Elaboración propia.

4.3. Presentación de resultados

Se consideró la interpretación y análisis de resultados según la comparación con los antecedentes y se precisaron las limitaciones generadas. Asimismo, las actividades desarrolladas para el logro de los objetivos se ejecutaron por cada etapa del modelo optimizado y de esta manera generar resultados de las evaluaciones realizadas. En ese marco, se consideraron cuatro (4) etapas para lograr los resultados: Preprocesamiento, evaluación del performance de los algoritmos, la correlación de los factores relevantes y la evaluación con métricas de performance del modelo de aprendizaje automático.

Respecto al análisis de resultados del preprocesamiento se identificó que el conjunto de datos (línea base) antes del preprocesamiento tenía 72% de celdas consistentes y luego del preprocesamiento se obtuvo el 100% de consistencia para un número total de 7,959 celdas. Asimismo, las técnicas realizadas para el preprocesamiento fue la de limpieza, normalización y categorización. Al respecto, se precisa que los resultados presentan similitud con los resultados obtenidos por Calabrese (2023) quien en su estudio consideró la etapa de limpieza de datos para evidenciar el mejoramiento del porcentaje de consistencia (incremento de 82% a 100%). Del mismo modo, los autores Lee et al. (2019) en su estudio precisaron que utilizaron un proceso de consistencia que mejoró la calidad en un 35% (incremento del 65% al 100%). Resultados similares fueron presentados por los autores Yu et al. (2019) quienes consideraron un híbrido entre la metodología KDD y la técnica de exploración de datos, donde la etapa de preprocesamiento fue la más determinante para generar confianza en los resultados, además precisaron que es necesario obtener un valor mayor o igual 95% de consistencia. Estos resultados evidenciaron que es importante la etapa de preprocesamiento para mejorar la consistencia del conjunto de datos.

Respecto a los resultados de la evaluación del performance de los algoritmos se determinaron que los algoritmos más idóneos fueron: XGBoost, SVM, LSTM, regresión logística, bosques aleatorios y red neuronal. Para el algoritmo de XGBoost se obtuvieron los resultados: precisión de 88%, exactitud de 86%, sensibilidad de 92% y f1-score de 90%, para el algoritmo de SVM se obtuvieron los resultados: precisión de 79%, exactitud de 78%, sensibilidad de 85% y f1-score de 84%, para el algoritmo de LSTM se obtuvieron los resultados: precisión de 75%, exactitud de 74%, sensibilidad de 82% y f1-score de 84%, para el algoritmo de regresión logística se obtuvieron los resultados: precisión de 85%, exactitud de 83%, sensibilidad de 95% y f1-score de 90%, para el algoritmo de bosques aleatorios se obtuvieron los resultados: precisión de 83%, exactitud de 83%, sensibilidad de 96% y f1-score de 90%. Para el algoritmo red neuronal se obtuvieron los resultados: precisión de 90%, exactitud de 87%, sensibilidad de 97% y f1-score de 92%. Los resultados obtenidos tienen similar proceso de evaluación realizado por el autor Calabrese (2023) que en su estudio identificó algoritmos de

aprendizaje automático para el cumplimiento de las metas; motivo por el cual, los algoritmos evaluados fueron: red neuronal, árboles de decisión, regresión lineal y Naïve Bayes, y las métricas fueron: precisión, exactitud, sensibilidad y f1-score, los cuales generaron resultados entre 85% a 92%. Del mismo modo, los autores Lee et al. (2019) en su estudio realizaron la evaluación de los algoritmos SVM, red neuronal y regresión lineal, donde obtuvieron resultados entre 92.7% y 94.5% referidas a las métricas evaluadas (precisión, exactitud, sensibilidad y f1-score). Se determinó que la etapa de evaluación del performance de los algoritmos es importante para la selección del mejor algoritmo y de esta manera generar métricas con resultados idóneos.

Respecto al análisis de la correlación de los factores relevantes se evidenció que las variables relevantes del conjunto de datos fueron determinadas por el cálculo realizado por sentencias codificadas usando la librería Seaborn de Python donde se generó una matriz de correlación de variables que permitió determinar las variables más relevantes y la relación entre ellas según el porcentaje obtenido luego del cálculo de correlación. De esa manera se obtuvo las variables con los valores más cercanos al 100%, los cuales fueron: Meses_A, Capital, Satisfacción_E, Rendimiento_Pago, Rentabilidad e Insolvencia, y las variables con valores entre 60% a 80% fueron: Acceso_Crédito y Letras_por_pagar. Los resultados fueron obtenidos en base al proceso de ingeniería de características donde se realizaron actividades que incluyeron técnicas especializadas de Shap y Lime para obtener los factores relevantes y la actividad de reducción de dimensionalidad realizada con la técnica PCA. Asimismo, se generaron nuevas características (índices) a partir de los factores relevantes: Riesgo, Confiabilidad, Sostenibilidad y Solidez, los cuales permitieron complementar los resultados del modelo. En ese marco, los autores Zhang et al. (2022) en su estudio identificaron cuatro (4) variables relevantes dentro de su modelo de predicción del riesgo de crédito, los cuales fueron: Solvencia, monto de crédito, índice de morosidad e índice de pago. Del mismo modo, los autores Bussmann et al. (2021) en su estudio precisaron que las variables más usadas en su conjunto de datos fueron: el índice de riesgo e índice de crédito, con ello lograron evaluar el riesgo de crédito en las empresas y generar correlaciones entre esas variables.

Respecto a los resultados obtenidos por las métricas de performance del modelo de aprendizaje automático se precisa que el algoritmo de red neuronal optimizado (MO2A) superó las métricas de los otros algoritmos considerados en la presente investigación, donde se obtuvo una precisión de 97%, exactitud de 97%, sensibilidad de 99% y f1-score de 98%. Asimismo, los resultados tienen coincidencias con el estudio realizado por Calabrese (2023) que en su estudio obtuvo 4 principales métricas: precisión (85%), exactitud (90%), sensibilidad (92%) y f1-score (89%). Estas métricas permitieron evaluar el riesgo crediticio y de quiebra, el cual fue diseñado específicamente para las MYPES y que permitió identificar empresas que necesitaban crédito, pero tenían ciertas condiciones desfavorables con el tiempo. La idea fundamental fue brindar soporte a las MYPES identificadas con riesgo crediticio y de quiebra. Del mismo modo, se obtuvo resultados similares al estudio de los autores Lee et al. (2019), quienes propusieron un modelo con base a las características de red neuronal para la predicción de precios de las acciones. Asimismo, consideraron el uso de índices económicos y financieros como predictores. Consideraron 4 principales métricas: precisión (94.5%), exactitud (92.7%), sensibilidad (93.5%) y f1-score (94%). Los cuales guardan resultados similares con el objetivo para predecir el desempeño financiero de las empresas utilizando una red neuronal artificial con algoritmos de retropropagación. Estos resultados permitieron identificar empresas que requerían invertir en acciones y no tenían sustento técnico para ello, la red neuronal evolucionó con la incorporación de más variables.

Respecto al objetivo general en la investigación se determinó desarrollar un modelo de aprendizaje automático para optimizar la predicción de la quiebra de empresas, donde se utilizó como base las 3 etapas propuestas por autores Alsheref et al. (2022): exploración de datos, selección del algoritmo e implementación. En ese sentido, se incorporaron dos actividades adicionales: Optimización del modelo y Evaluación del performance del modelo. Todo ello, permitió que los resultados de las métricas generen resultados más idóneos y verídicos. Estos resultados tienen similar comportamiento con lo precisado en el estudio propuesto por los autores Zhang et al. (2022) quienes precisaron la importancia de incorporar el índice financiero en un modelo de bosque aleatorio mejorado mediante las etapas

definidas con la metodología KDD, donde consideraron el mejoramiento del modelo en base al entrenamiento respectivo, todo ello permitió predecir los riesgos asociados al crédito y a la quiebra de MYPES. Ese sentido, obtuvieron las métricas: precisión (88%), exactitud (89%), sensibilidad (90%) y f1-score (88%).

Resultados similares se presentaron en el estudio propuesto por los autores Bussmann et al. (2021), donde se precisó el tratamiento de cinco modelos de redes neuronales con características de predicción y asociados a los riesgos de quiebra, crédito y rendimiento de pago. Se demostró que la red neuronal profunda desempeña con mayor precisión en la predicción del riesgo crediticio que otros modelos. La propuesta fue un modelo desarrollado mediante la metodología KDD con aprendizaje profundo para mejorar la gestión de riesgos financieros y los experimentos con datos comerciales. Consideraron 4 principales métricas: precisión (94%), exactitud (94%), sensibilidad (93%) y f1-score (94%). En este estudio se demostró la superioridad del aprendizaje profundo sobre otras técnicas de aprendizaje automático. Resultados similares se precisaron en el estudio de los autores Balcaen y Ooghe (2019) donde obtuvieron métricas de precisión (93%), exactitud (92%), sensibilidad (92%) y f1-score (93%) que permitió obtener el objetivo general de predecir el riesgo crediticio y de quiebra de las MYPES en China y que este enfoque ha mejorado significativamente la predicción en un modelo de red neuronal. Asimismo, los resultados obtenidos tienen similitud con el trabajo realizado por los autores Cantarero et al. (2019) donde lograron realizar pronósticos más acertados en la predicción de riesgo de quiebra y crediticio de las MYPES mediante un modelo desarrollado con las etapas de la metodología CRISP-DM, donde utilizaron un algoritmo de regresión logística para lograr el objetivo propuesto. En ese sentido, los resultados de las métricas fueron mayores al 90%: precisión (90%), exactitud (91%), sensibilidad (90%) y f1-score (90%). Propiciando así valores con una significancia alta para aprobar la hipótesis planteada. Del mismo modo, los autores Ravisankar et al. (2020), utilizaron un modelo similar de regresión logística, donde se enfocaron en utilizar variables de razón financiera y datos de sentencias judiciales de MYPES Chinas. Asimismo, consideraron el objetivo general de predecir el riesgo de quiebra empresarial de las MYPE y obtuvieron valores significativos que ayudaron a comprobar la hipótesis planteada.

Por otro lado, los autores Han y Gao (2019), en su estudio obtuvieron resultados idóneos para conseguir el objetivo planteado con un modelo desarrollado con la metodología KDD en base a un algoritmo de bosque aleatorio mejorado mediante el uso de la relación financiera y los datos de transacciones para predecir el riesgo de quiebra de las MYPES en Gran Bretaña, los cuales lograron generar valores significativos para comprobar su hipótesis planteada. Resultados similares se obtuvieron en el estudio de los autores Kuster et al. (2023), donde propusieron de un sistema de diagnóstico temprano financiero para evaluar la insolvencia de las empresas mediante una red neuronal artificial, considerando las métricas de precisión (91%), exactitud (90%), sensibilidad (89%) y f1-score (91%), y el método de exploración de datos. La obtención de resultados mediante la variable de insolvencia mejoró la predicción de la quiebra de empresas Serbias. En relación a estos resultados se obtuvo similitud respecto al estudio de los autores Yu et al. (2019), donde propusieron un modelo de fusión adaptativo entre metodologías KDD y de exploración de datos para predecir el riesgo de quiebra y consideraron 4 principales métricas: precisión (82%), exactitud (83%), sensibilidad (81%) y f1-score (81%), lo que significa que ante la incorporación de mayor número de variables y correlaciones se obtendrá mejores resultados sobre lo que se requiere pronosticar.

En la investigación se presentaron detalles que propiciaron limitaciones en la obtención de datos reales de las empresas encuestadas, muchas veces el Gerente o Tutor de la empresa desconocía sobre los indicadores financieros requeridos para el presente estudio. En esa casuística se consideró brindar asistencia técnica y un trato más personalizado. Asimismo, se consideró un margen de error de $\pm 5\%$, el cual significa que algunos datos fueron registrados de manera subjetiva.

V. CONCLUSIONES

- 5.1. Los resultados del preprocesamiento identificaron que el conjunto de datos (línea base) antes del preprocesamiento tenía 72% de celdas consistentes y luego del preprocesamiento el 100% de consistencia para un número total de 7,959 celdas. Asimismo, las técnicas realizadas para el preprocesamiento fueron limpieza, normalización y categorización. Para la comprobación de la hipótesis se utilizó la prueba Wilcoxon donde se obtuvo 0.045 de significancia (valor menor a 0.05). En ese sentido, se aprobó la hipótesis formulada y se concluye que se presentaron diferencias significativas antes y después del preprocesamiento.
- 5.2. Los resultados obtenidos determinaron que la etapa de evaluación del performance de los algoritmos es importante para la selección del mejor algoritmo. Se determinó que los algoritmos más idóneos fueron: XGBoost, SVM, LSTM, regresión logística, bosques aleatorios y red neuronal. En ese sentido, los algoritmos que obtuvieron un mejor nivel de concordancia según la medición Kappa de Cohen fueron: Regresión logística y red neuronal con >0.81 , donde el algoritmo red neuronal obtuvo un valor mayor 0.88. Asimismo, se obtuvo la significancia menor a 0.05, el cual permitió aprobar la hipótesis formulada.
- 5.3. En el análisis de la correlación se determinaron los factores relevantes del modelo mediante el cálculo de la correlación variables, donde se consideró los valores más cercanos al 100%. En ese contexto, se identificaron las variables: Rendimiento de pago, Capital, Rentabilidad, Tiempo de actividad, Satisfacción Empresarial e Insolvencia. Adicionalmente a ello, se consideraron variables que se encontraron entre el 60% al 80% como variables de complemento: Acceso a crédito y letras por pagar. Se obtuvo que

las correlaciones realizadas con la medida Rho Spearman determinaron resultados de significancia menores al 0.05. Motivo por el cual, se aprobó la hipótesis formulada.

- 5.4. En la etapa de resultados obtenidos por las métricas de performance se identificó que el algoritmo de red neuronal propuesto en el modelo optimizado de aprendizaje automático (MO2A) superó las métricas de los otros algoritmos, obteniéndose una precisión de 97%, exactitud de 97%, sensibilidad de 99% y f1-score de 98%. Se identificó 134 (35.36%) empresas que se encuentran en riesgo de quiebra del total de 379 (100%). Asimismo, obtuvo un valor mayor 0.91 de concordancia con la medida Kappa de Cohen y significancia menor a 0.05. En ese sentido, la hipótesis se aprobó y validó con este algoritmo.
- 5.5. Los objetivos específicos se realizaron mediante actividades que permitieron ejecutar cada etapa del modelo optimizado y de esta manera generar resultados. En ese marco, se consideraron cuatro (4) etapas: Preprocesamiento, evaluación del performance de los algoritmos, la correlación de los factores relevantes y la evaluación con métricas de performance del modelo. Se ha demostrado que los resultados obtenidos fueron más eficientes y efectivos para determinar la predicción de la quiebra de empresas. De esta manera, la hipótesis general fue comprobada de manera significativa.

VI. RECOMENDACIONES

- 6.1. Desarrollar procedimientos de comparación y evaluación de algoritmos diversos de aprendizaje automático y otros tipos de tecnología. Los cuales permitirán predecir, clasificar y evaluar mediante métricas precisas y elaboradas con idoneidad para su tratamiento.
- 6.2. Gestionar la identificación, análisis y desarrollo de metodologías tradicionales, combinadas o híbridas que permitan comparar resultados y generar enfoques con escenarios propicios sobre el contexto propio. Todo ello es necesario para el desarrollo de soluciones con resultados más idóneos y específicos.
- 6.3. Desarrollar un sistema inteligente que automatice los procesos de predicción de la quiebra de empresas. De esta manera, el proceso de implementación del modelo y la utilización de métricas con un adecuado entrenamiento permitiría obtener resultados inmediatos y específicos.
- 6.4. Desarrollar estudios alternativos que propicien el desarrollo de evaluaciones y la comparación de resultados en contextos comunes o dispersos mediante el mejoramiento de modelos y la optimización de procesos para escenarios alternativos o con pronósticos complejos. Todo ello permitirá la generación de conocimiento para brindar soluciones innovadoras y en tiempos eficientes.
- 6.5. Para trabajos futuros se debe priorizar la incorporación de variables que representen aspectos significativos de la empresa y del mercado como factores internos y externos. Asimismo, considerar el estudio de correlaciones para determinar los factores que tienen una mayor o menor relación entre las variables. De esta manera, el modelo obtendrá resultados más cercados al 100% y propiciarán una mejora continua.

REFERENCIAS BIBLIOGRÁFICAS

- Alsheref, F. K., Fattoh, I. E., & M. Ead, W. (2022). Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms. Computational Intelligence and Neuroscience, 2022, 1-9. <https://doi.org/10.1155/2022/7728668>
- Balcaen, S., & Ooghe, H. (2019). Alternative methodologies in studies on business failure: do they produce better results than the classical statistical methods? Faculteit Economie en Bedrijfskunde. Hoveniersberg. Gante. Bélgica. https://www.researchgate.net/publication/23646614_The_Ooghe-Joos-De_Vos_Failure_Prediction_Models_A_Cross-Industry_Validation
- Beck S., Kuner M., Haar M. Daubmann A., Semmann M., & Kluge S. (2024). Evaluating the accuracy and reliability of AI chatbots in disseminating the content of current resuscitation guidelines: a comparative analysis between the ERC 2021 guidelines. Scand J Trauma Resusc Emerg Med, Volume 32, article number 95. <https://doi.org/10.1186/s13049-024-01266-2>
- Ben-Sassi, D. (2022). A rough set-based Competitive Intelligence approach for anticipating competitor's action. Canadá. <https://doi.org/10.1016/j.eswa.2022.117523>
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. Computational Economics, 57(1), 203-216. <https://doi.org/10.1007/s10614-020-10042-0>
- Calabrese, R. (2023). Contagion effects of UK small business failures: A spatial hierarchical autoregressive model for binary data. European Journal of Operational Research, 305(2), 989-997. <https://doi.org/10.1016/j.ejor.2022.06.027>
- Cantarero, S., Gonzales Loureiro, M., & Puig, F. (2019). Efectos de la crisis económica sobre el emprendimiento en empresas de economía social en

- España: un análisis espacial. Madrid, España: Universidad Complutense de Madrid. <https://www.redalyc.org/articulo.oa?id=36754074002>
- Chávez-Zavaleta, R., Medina-Ordoñez, L., Ochoa-Cuartas, B., Palomino-Tiznado, M., & Cáceda-Fárfan, A. (2024). Artificial neural network for finances in MYPES according to internal factors in response to the demand of the Chancay Mega port. <https://doi.org/10.18687/LACCEI2024.1.1.1569>
- Deichmann, U. (2024). Contrasting philosophical and scientific views in the long history of studying the generation of form in development. *BioSystems*. Volume 242. <https://doi.org/10.1016/j.biosystems.2024.105260>
- Gao, B., & Balyan, V. (2022). Construction of a financial default risk prediction model based on the LightGBM algorithm. *Journal of Intelligent Systems*, 31(1), 767-779. <https://doi.org/10.1515/jisys-2022-0036>
- Garcete, A., Benítez, R., Pinto-Roa, D., & Vazquez, A. (2019). Técnica de pronóstico de la demanda basada en Business Intelligence y Machine Learning. Simposio Argentino Sobre Tecnología y Sociedad, 193–202. http://sedici.unlp.edu.ar/bitstream/handle/10915/64728/Documento_completo.pdf-PDFA.pdf?sequence=1
- Han, X., & Gao, S. (2019). A Chain Multiple Mediation Model Linking Strategic, Management, and Technological Innovations to Firm Competitiveness. P. R. China: Xi'an Jiaotong University. <https://doi.org/10.7819/rbgn.v21i5.4030>
- Hernández-Sampieri, R., Fernández C., & Baptista P. (2018). Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta, Ciudad de México, México: Editorial Mc Graw Hill Education, Año de edición: 2018, ISBN: 978-1-4562-6096-5, 714 p. <https://doi.org/10.22201/fesc.20072236e.2019.10.18.6>
- Horváthová, J., Mokrišová, M. & Schneider, A. (2024). The Application of Machine Learning in Diagnosing the Financial Health and Performance of Companies in the Construction Industry. University of Prešov, Konštantínova, Slovakia. <https://doi.org/10.3390/info15060355>
- Kanzari, A., Rasmussen, J., Nehler, H., & Ingelsson, F. (2022). How financial performance is addressed in light of the transition to circular business models—A systematic literature review. *Journal of Cleaner Production*, 376, 134134. <https://doi.org/10.1016/j.jclepro.2022.134134>

- Kuster, D., Vukovic, B., Milutinovic, S., Peštovic, K., Tica, T. & Jakšić. (2023). Early Insolvency Prediction as a Key for Sustainable Business Growth. University of Novi Sad, Serbia. <https://doi.org/10.3390/su152115304>
- Lee, J., Jang, D., & Park, S. (2019). Deep Learning-Based Corporate Performance Prediction Model Considering Technical Capability. Sustainability, 9(6), 899. <https://doi.org/10.3390/su9060899>
- Mobbs A., Kahn M., Williams G., Mentiplay B., Pua Y., & Clark R. (2024). Mobbs et al. Journal of NeuroEngineering and Rehabilitation. <https://doi.org/10.1186/s12984-024-01406-w>
- Montoya, I., & Montoya, A. (2011). La conexión entre el pensamiento sistémico y nociones de oposiciones y triadas, desde la perspectiva compleja. Escuela Superior de Comercio y Administración, Unidad Santo Tomás Distrito Federal, México. <http://www.redalyc.org/articulo.oa?id=456045213003>
- Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2023). On the combination of graph data for assessing thin-file borrowers' creditworthiness. Expert Systems with Applications, 213, 118809. <https://doi.org/10.1016/j.eswa.2022.118809>
- Muthukumaran, K., & Hariharanath, K. (2022). Deep Learning Enabled Financial Crisis Prediction Model for Small-Medium Sized Industries. India. <https://doi.org/10.32604/iasc.2023.025968>
- Nicodemo C., & Satorra, A. (2022). Exploratory data analysis on large data sets: The example of salary variation in Spanish Social Security Data. Business Research Quarterly 2022, Vol. 25(3) 83-294. DOI: <https://doi.org/10.1177/2340944420957335>
- Ñaupas P. (2018), Metodología de la investigación Cuantitativa - Cualitativa y Redacción de la Tesis. Colombia: Editorial Ediciones de la U. <https://corladancash.com/wp-content/uploads/2020/01/Metodologia-de-la-inv-cuanti-y-cuali-Humberto-Naupas-Paitan.pdf>
- Pandey B., Veeramanickam M., Ahmad, S., Rodriguez C., & Esenarro, D., (2023). ExpSSOA-Deep Maxout: Exponential Shuffled shepherd optimization based Deep maxout network for intrusion detection using big data in cloud computing framework, Computers & Security, Vol. 124, 102975, ISSN0167-4048, <https://doi.org/10.1016/j.cose.2022.102975>

- Pérez-Escobar, J. A. & Sarikaya D. (2024). Philosophical Investigations into AI Alignment: A Wittgensteinian Framework. *Philosophy & Technology*. 37:80 <https://doi.org/10.1007/s13347-024-00761-9>
- Pratt, M., Boudhane, M., & Cakula, S. (2021). Employee Attrition Estimation Using Random Forest Algorithm. *Baltic Journal of Modern Computing*, 9(1). <https://doi.org/10.22364/bjmc.2021.9.1.04>
- Ramos-Galarza C. (2021). Diseños de investigación experimental. *CienciAmérica*, vol. 10, no. 1, pp. 1–7, Feb. 2021, Doi: <http://dx.doi.org/10.33210/ca.v10i1.356>
- Ravisankar, P., Ravi, V., Raghava, G., & Bose, I. (2020). Detection of financial statement fraud and feature selection using data mining techniques. Institute for Development and Research in Banking Technology, Hyderabad, India. <https://doi.org/10.1016/j.dss.2010.11.006>
- Raza, A., Munir, K., Almutairi, M., Younas, F. & Fareed, M. (2022). Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences*, 12(13), 6424. <https://doi.org/10.3390/app12136424>
- Reyes Meza, C. & Sánchez Carlessi, H. (2015). Metodología y diseños en la investigación científica. Lima, Perú: Visión Universitaria. <https://isbn.cloud/9786124684227/metodologia-y-diseno-en-la-investigacion-cientifica/>
- Sulastri, R., & Janssen, M. (2022). The elements of the Peer-to-peer (P2P) lending system: A Systematic Literature Review. 15th International Conference on Theory and Practice of Electronic Governance, 424-431. <https://doi.org/10.1145/3560107.3560172>
- Trujillo, D. (2017). Aplicación de metodologías Aprendizaje automático en la gestión de riesgo de crédito. <https://doi.org/10.26439/interfases2022.n015.5898>
- Wang, X., Yang, J., Liu, Y. & Wang Y. (2024). Parallel intelligence in three decades: a historical review and future perspective on ACP and cyber-physical-social systems. *Artificial Intelligence Review* (2024) 57:255. <https://doi.org/10.1007/s10462-024-10861-9>
- Xiaolei Y., Zhixin X., Junhui S. & Yongshan L. (2024). Credit risk prediction for small and micro enterprises based on federated transfer learning frozen network parameters. <https://doi.org/10.1016/j.inca.2024.104009>

- Yahia, NB, Hlel, J., & Colomo-Palacios, R. (2021). From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. *IEEE Access*, 9, 60447-60458. <https://doi.org/10.1109/ACCESS.2021.3074559>
- Yao, G., Hu, X., & Wang, G. (2022). A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain. *Expert Systems with Applications*, 200, 117002. <https://doi.org/10.1016/j.eswa.2022.117002>
- Yu, Q., Lendasse, A., & Severin, E. (2019). Ensemble KNNs for Bankruptcy Prediction. Helsinki University of Technology. Computer Science Department. https://www.researchgate.net/publication/255669581_Ensemble_KNNs_for_Bankruptcy_Prediction
- Zhang, W., Yan, S., Li, J., Tian, X., & Yoshida, T. (2022). Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. *Transportation Research Part E: Logistics and Transportation Review*, 158, 102611. <https://doi.org/10.1016/j.tre.2022.102611>

ANEXOS

INSTRUMENTO (CUESTIONARIO) PARA LA RECOLECCIÓN DE DATOS

Nº	
Encuesta:	
Fecha:	

Estimado (a), este cuestionario tiene el objetivo de recopilar datos que permitan generar una matriz de conjuntos de datos sobre aspectos del entorno empresarial y la situación financiera de su empresa. Agradezco tu participación.

INSTRUCCIONES: Registre la respuesta en el casillero que corresponde.

Nº	Dimensión	Preguntas	Respuesta	Intervalo
1	Entorno Empresarial	¿Cuántos meses de actividad tiene la empresa?		[0-36]
2		¿La empresa tiene planes normativos?		[0-1]
3		¿Cuántos empleados tiene trabajando en la empresa?		[0-500]
4		¿La empresa tiene sistemas de información?		[0-1]
5		¿Cuánto considera que es la satisfacción empresarial / clima laboral?		[0-5]
6		¿Tiene acceso a mercados externos (nacionales e internacionales)?		[0-1]
7		¿Conforma alguna sociedad?		[0-1]
8		¿Cuántos competidores directos tiene la empresa?		[0-10]
9		¿La empresa experimentó siniestros?		[0-1]
10		¿La empresa ha tenido imprevistos?		[0-1]
11	Situación Financiera	¿Cuál es la capacidad de endeudamiento?		[0-5]
12		¿Cuánto fue el capital inicial de la empresa?		[0-999999]
13		¿Tiene acceso a crédito?		[0-1]
14		¿Tiene letras pendientes de pago?		[0-1]
15		¿Cuál es la calificación crediticia?		[0-5]
16		¿Cuál es el rendimiento de pago?		[0-5]
17		¿Cuál es la tasa de utilidades?		[0-100%]
18		¿Cuál es el índice de rentabilidad?		[0-5]
19		¿Cuál es índice de eficiencia en ventas?		[0-5]
20		¿Cuál es índice de insolvencia?		[0-5]

VALIDACIÓN DE CONTENIDO DEL INSTRUMENTO (CUESTIONARIO)

INSTRUCCIÓN: A continuación, se le hace llegar el presente cuestionario que permitirá recoger los datos para la investigación: Modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas. Por lo que se le solicita que tenga a bien evaluar este instrumento. Si tiene alguna sugerencia, observación o comentario puede registrarlas (si lo considera). De esta manera, realizar las correcciones pertinentes. Los criterios de validación de contenido son:

Criterios	Detalle	Calificación
Suficiencia	El ítem pertenece a la dimensión y basta para obtener la medición de esta	1: de acuerdo 0: en desacuerdo
Claridad	El ítem se comprende fácilmente, es decir, su sintáctica y semántica son adecuadas	1: de acuerdo 0: en desacuerdo
Coherencia	El ítem tiene relación lógica con el indicador que está midiendo	1: de acuerdo 0: en desacuerdo
Relevancia	El ítem es esencial o importante, es decir, debe ser incluido	1: de acuerdo 0: en desacuerdo

Nota. Criterios adaptados de la propuesta de Escobar y Cuervo (2008).

MATRIZ DE VALIDACIÓN DEL INSTRUMENTO (CUESTIONARIO)

Variable independiente: El aprendizaje automático es una categoría de la inteligencia artificial que considera el propósito de generar un aprendizaje más eficiente y efectivo mediante instrucciones precisas y complejas. Los autores Chávez-Zavaleta et al. (2024), consideraron que el aprendizaje tradicional tiene muchas variables, pero todas se enfocan en procesos sin un seguimiento adecuado o actividades que permitan crecer y mejorar.

Variable dependiente: Quiebra de empresas. Ravisankar et al. (2020) consideraron que la quiebra empresarial es una etapa trascendente para una institución y que genera diversas actividades definitorias en un panorama de infortunio y desgaste.

Dimensión	Ítem o enunciado	Suficiencia	Claridad	Coherencia	Relevancia	Observación
Entorno Empresarial	¿Cuántos meses de actividad tiene la empresa?	1	1	1	1	Ítem claro y relevante
	¿La empresa tiene planes normativos?	1	1	1	1	Ítem claro y relevante
	¿Cuántos empleados tiene trabajando en la empresa?	1	1	1	1	Ítem claro y relevante
	¿La empresa tiene sistemas de información?	1	1	1	1	Ítem claro y relevante

	¿Cuánto considera que es la satisfacción empresarial / clima laboral?	1	1	1	1	Ítem claro y relevante
	¿Tiene acceso a mercados externos (nacionales e internacionales)?	1	1	1	1	Ítem claro y relevante
	¿Conforma alguna sociedad?	1	1	1	1	Ítem claro y relevante
	¿Cuántos competidores directos tiene la empresa?	1	1	1	1	Ítem claro y relevante
	¿La empresa experimentó siniestros?	1	1	1	1	Ítem claro y relevante
	¿La empresa ha tenido imprevistos?	1	1	1	1	Ítem claro y relevante
Situación Financiera	¿Cuál es la capacidad de endeudamiento?	1	1	1	1	Ítem claro y relevante
	¿Cuánto fue el capital inicial de la empresa?	1	1	1	1	Ítem claro y relevante
	¿Tiene acceso a crédito?	1	1	1	1	Ítem claro y relevante
	¿Tiene letras pendientes de pago?	1	1	1	1	Ítem claro y relevante
	¿Cuál es la calificación crediticia?	1	1	1	1	Ítem claro y relevante
	¿Cuál es el rendimiento de pago?	1	1	1	1	Ítem claro y relevante
	¿Cuál es la tasa de utilidades?	1	1	1	1	Ítem claro y relevante
	¿Cuál es el índice de rentabilidad?	1	1	1	1	Ítem claro y relevante
	¿Cuál es índice de eficiencia en ventas?	1	1	1	1	Ítem claro y relevante
	¿Cuál es índice de insolvencia?	1	1	1	1	Ítem claro y relevante



Dr. Alfredo Daza Vergaray
DNI 40466240



Dr. Javier Arturo Gamboa Cruzado
DNI 17906323



Willy Gustavo Ugarte Rojas
DNI 43673615

INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO DE REGRESIÓN LOGÍSTICA

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Test
Investigador	Dino Michael Quinteros Navarro
Fecha de inicio	03/05/2024
Algoritmo	REGRESIÓN LOGÍSTICA

MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	Positivo (P)
Real	Negativo	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	Indicador	Medida	Fórmula	Precisión
1	Precisión	Razón	$(TP/(TP+FP)) * 100$	
2	Sensibilidad	Razón	$(TP/(TP+FN)) * 100$	
3	Exactitud	Razón	$(TP+TN)/(TP+TN+FP+FN)$	
4	F1-score	Razón	$2 * ((\text{Precisión} * \text{Sensibilidad}) / (\text{Precisión} + \text{Sensibilidad}))$	



Dr. Alfredo Daza Vergaray
DNI 40466240



Dr. Javier Arturo Gamboa Cruzado
DNI 17906323



Willy Gustavo Ugarte Rojas
DNI 43673615

INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO BOSQUES ALEATORIOS

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Test
Investigador	Dino Michael Quinteros Navarro
Fecha de inicio	03/05/2024
Algoritmo	BOSQUES ALEATORIOS

MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	Positivo (P)
Real	Negativo	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	Indicador	Medida	Fórmula	Precisión
1	Precisión	Razón	$(TP/(TP+FP)) * 100$	
2	Sensibilidad	Razón	$(TP/(TP+FN)) * 100$	
3	Exactitud	Razón	$(TP+TN)/(TP+TN+FP+FN)$	
4	F1-score	Razón	$2 * ((\text{Precisión} * \text{Sensibilidad}) / (\text{Precisión} + \text{Sensibilidad}))$	



Dr. Alfredo Daza Vergaray
DNI 40466240



Dr. Javier Arturo Gamboa Cruzado
DNI 17906323



Willy Gustavo Ugarte Rojas
DNI 43673615

INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO RED NEURONAL

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Test
Investigador	Dino Michael Quinteros Navarro
Fecha de inicio	03/05/2024
Algoritmo	Red Neuronal

MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	Positivo (P)
Real	Negativo	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	Indicador	Medida	Fórmula	Precisión
1	Precisión	Razón	$(TP/(TP+FP)) * 100$	
2	Sensibilidad	Razón	$(TP/(TP+FN)) * 100$	
3	Exactitud	Razón	$(TP+TN)/(TP+TN+FP+FN)$	
4	F1-score	Razón	$2 * ((\text{Precisión} * \text{Sensibilidad}) / (\text{Precisión} + \text{Sensibilidad}))$	



Dr. Alfredo Daza Vergaray
DNI 40466240



Dr. Javier Arturo Gamboa Cruzado
DNI 17906323



Willy Gustavo Ugarte Rojas
DNI 43673615

VALIDACIÓN DEL INSTRUMENTO

I. DATOS GENERALES:

Apellidos y Nombres del experto: Daza Vergaray, Alfredo

Título y/o grado: Dr. Ingeniería de Sistemas

Fecha: 03/05/2024

Instrumento: Cuestionario

Autores: Dino Michael Quinteros Navarro

Título de la investigación:

Modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas.

II. ASPECTOS DE VALIDACIÓN:

INDICADORES	CRITERIOS	Deficiente 0 – 20%	Regular 21 – 50%	Bueno 51 – 70%	Muy bueno 71 – 80%	Excelente 81 – 100%
1. Claridad	Está formulado con el lenguaje apropiado.					X
2. Objetividad	Está expresado en conducta observable.					X
3. Actualidad	Es adecuado al avance de la ciencia.					X
4. Organización	Existe una organización lógica.					X
5. Suficiencia	Comprende los aspectos de cantidad y calidad.					X
6. Intencionalidad	Es adecuado para valorar aspectos del sistema metodológico y científico.					X
7. Consistencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					X
8. Coherencia	Entre los índices, indicadores, dimensiones.					X
9. Metodología	Responde al propósito del trabajo bajo los objetivos a lograr.					X
10. Pertinencia	El instrumento es adecuado al tipo de investigación.					X
Promedio de Validación						90 %

III. **Promedio de Valoración: 90%**

IV. **Observaciones:**



Dr. Alfredo Daza Vergaray
DNI 40466240

VALIDACIÓN DEL INSTRUMENTO

I. DATOS GENERALES:

Apellidos y Nombres del experto: Gamboa Cruzado, Javier Arturo

Título y/o grado: Dr. Ingeniería de Sistemas

Fecha: 03/05/2024

Instrumento: Cuestionario

Autores: Dino Michael Quinteros Navarro

Título de la investigación:

Modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas

II. ASPECTOS DE VALIDACIÓN:

INDICADORES	CRITERIOS	Deficiente 0 – 20%	Regular 21 – 50%	Bueno 51 – 70%	Muy bueno 71 – 80%	Excelente 81 – 100%
1. Claridad	Está formulado con el lenguaje apropiado.					X
2. Objetividad	Está expresado en conducta observable.					X
3. Actualidad	Es adecuado al avance de la ciencia.					X
4. Organización	Existe una organización lógica.					X
5. Suficiencia	Comprende los aspectos de cantidad y calidad.					X
6. Intencionalidad	Es adecuado para valorar aspectos del sistema metodológico y científico.					X
7. Consistencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					X
8. Coherencia	Entre los índices, indicadores, dimensiones.					X
9. Metodología	Responde al propósito del trabajo bajo los objetivos a lograr.					X
10. Pertinencia	El instrumento es adecuado al tipo de investigación.					X
Promedio de Validación						90 %

III. **Promedio de Valoración: 90%**

IV. **Observaciones:**



Dr. Javier Arturo Gamboa Cruzado
DNI 17906323

VALIDACIÓN DEL INSTRUMENTO

I. DATOS GENERALES:

Apellidos y Nombres del experto: Ugarte Rojas, Willy Gustavo

Título y/o grado: Dr. Ingeniería de Sistemas

Fecha: 03/05/2024

Instrumento: Cuestionario

Autor: Dino Michael Quinteros Navarro

Título de la investigación:

Modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas

II. ASPECTOS DE VALIDACIÓN:

INDICADORES	CRITERIOS	Deficiente 0 – 20%	Regular 21 – 50%	Bueno 51 – 70%	Muy bueno 71 – 80%	Excelente 81 – 100%
1. Claridad	Está formulado con el lenguaje apropiado.					X
2. Objetividad	Está expresado en conducta observable.					X
3. Actualidad	Es adecuado al avance de la ciencia.					X
4. Organización	Existe una organización lógica.					X
5. Suficiencia	Comprende los aspectos de cantidad y calidad.					X
6. Intencionalidad	Es adecuado para valorar aspectos del sistema metodológico y científico.					X
7. Consistencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					X
8. Coherencia	Entre los índices, indicadores, dimensiones.					X
9. Metodología	Responde al propósito del trabajo bajo los objetivos a lograr.					X
10. Pertinencia	El instrumento es adecuado al tipo de investigación.					X
Promedio de Validación						90 %

III. **Promedio de Valoración: 90%**

IV. **Observaciones:**



Willy Gustavo Ugarte Rojas
DNI 43673615

INSTRUMENTO: Cuestionario

INVESTIGADOR: Dino Michael Quinteros Navarro

TÍTULO DE LA INVESTIGACIÓN:

Modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas

PRUEBA FIABILIDAD: Alpha de Cronbach

FECHA: 03/05/2024

Escala: ALL VARIABLES

Resumen de procesamiento de casos

		N	%
Casos	Válido	20	100.00
	Excluido ^a	0	0.00
	Total	20	100.00

a. La eliminación por lista se basa en todas Variables del procedimiento

Alfa de Cronbach	N de elementos
0.817	20

DESCRIPCIÓN DEL PROCESO

El valor encontrado del índice Alfa de Cronbach de **0.817** se encuentra en el rango de confiabilidad, lo que garantiza la aplicabilidad del instrumento.



Dr. Alfredo Daza Vergaray
DNI 40466240



Dr. Javier Arturo Gamboa Cruzado
DNI 17906323



Willy Gustavo Ugarte Rojas
DNI 43673615

PRUEBA DE CONFIABILIDAD DEL INSTRUMENTO ALPHA DE CRONBACH

Sujeto	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20	Total
1	5	4	4	4	5	5	5	5	5	5	5	5	4	5	4	4	4	3	4	4	57
2	4	4	5	4	3	4	4	3	3	4	4	5	4	4	4	4	4	3	4	4	47
3	4	5	4	4	5	5	5	3	4	3	3	5	4	4	5	4	4	3	4	4	50
4	4	3	2	2	2	2	3	3	3	2	2	3	3	4	4	4	3	3	4	4	31
5	5	4	5	4	4	4	4	2	4	2	4	4	4	4	5	4	4	4	4	4	46
6	5	3	2	3	3	3	4	4	4	4	4	4	4	4	4	4	4	3	2	4	43
7	2	2	4	4	4	3	3	2	3	4	4	4	3	4	4	5	4	2	2	4	39
8	5	4	3	4	4	3	5	4	4	4	5	5	4	4	3	3	3	2	4	3	50
9	3	3	1	1	2	2	2	2	4	4	3	3	3	4	4	4	4	3	4	4	30
10	3	3	2	2	2	2	2	3	2	4	3	3	3	4	4	3	4	4	3	3	31
11	3	4	4	3	2	3	3	2	3	3	4	3	4	4	4	4	4	4	3	4	37
12	4	4	3	5	2	3	5	3	4	4	4	4	4	4	4	3	4	4	2	4	45
13	4	4	4	5	3	3	4	4	3	3	4	2	3	4	4	4	4	4	3	4	43
14	2	2	3	2	3	2	2	2	2	3	4	3	3	3	3	4	4	4	3	2	30
15	2	2	2	2	2	2	2	4	3	3	3	2	3	3	4	4	4	4	3	5	29
16	4	3	3	4	3	3	2	3	3	3	3	2	3	2	3	4	3	4	2	2	36
17	2	2	2	2	2	2	2	2	2	2	3	2	4	4	4	4	4	4	4	3	25
18	3	4	4	3	4	3	5	4	3	4	3	4	4	3	4	3	4	4	2	5	44
19	3	4	5	2	3	2	5	5	2	2	2	4	4	3	4	5	4	4	2	2	39
20	4	4	5	4	2	3	2	4	2	2	3	3	4	4	5	4	4	4	4	3	38
Varianzas	1.048	0.740	1.428	1.260	1.000	0.848	1.548	0.960	0.728	0.788	0.650	1.050	0.240	0.388	0.300	0.290	0.128	0.450	0.728	0.740	68.350

$k = 20$

$V_i = 15.308$

$V_t = 68.350$

$\alpha = 0.817$

MATRIZ CONSISTENCIA

TITULO	Modelo de aprendizaje automático para mejorar la predicción de la quiebra de empresas				
PROBLEM	OBJETIVOS	START OF THE ART	HYPOTHESYS, PROPOSAL	VARIABLES	METHODOLOGY
Problema principal: ¿De qué manera un modelo de aprendizaje automático optimizará la predicción de la quiebra de empresas?	Objetivo general: Desarrollar un modelo de aprendizaje automático para optimizar la predicción de la quiebra de empresas.	Calabrese (2023): Contagion effects of UK small business failures: A spatial hierarchical autoregressive model for binary data. Lee et al. (2019): Deep Learning-Based Corporate Performance Prediction Model Considering Technical Capability. Zhang et al. (2022): Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. Bussmann et al. (2021): Explainable Machine Learning in Credit Risk Management. Balcaen y Ooghe (2019): Alternative methodologies in studies on business failure: do they produce better results than the classical statistical methods? Faculteit Economie en Bedrijfskunde. Cantarero, S., Gonzales Loureiro, M. y Puig, F. (2019): Efectos de la crisis económica sobre el emprendimiento en empresas de economía social en España: un análisis espacial. Han, X. y Gao, S. (2019): A Chain Multiple Mediation Model Linking Strategic, Management, and Technological Innovations to Firm Competitiveness. Ravisankar, P., Ravi, V., Raghava, G., Bose, I. (2020): Detection of financial statement fraud and feature selection using data mining techniques.	Hipótesis general: El desarrollo de un modelo de aprendizaje automático optimiza la predicción de la quiebra de empresas.	Variable Independiente Modelo de aprendizaje automático	Tipo de investigación: Aplicada Diseño: Experimental Enfoque: Cuantitativo Método: Análisis Exploratorio de Datos (EDA) Población: 25,880 Muestra: 379
Problemas específicos: ¿Cómo preprocesar la data colectada de empresas para optimizar la predicción del modelo de aprendizaje automático? ¿Cómo evaluar el performance de los algoritmos de aprendizaje automático para optimizar la predicción del modelo de la quiebra de empresas? ¿Cómo correlacionar los factores relevantes de las empresas para optimizar el diseño del modelo de aprendizaje automático de la predicción de la quiebra? ¿Cómo evaluar con métricas de performance el modelo de aprendizaje automático para la predicción de la quiebra de empresas?	Objetivos específicos: Preprocesar la data colectada de empresas para optimizar la predicción del modelo de aprendizaje automático. Evaluar el performance de los algoritmos de aprendizaje automático para optimizar la predicción del modelo de la quiebra de empresas. Correlacionar los factores relevantes de las empresas para optimizar el diseño del modelo de aprendizaje automático de la predicción de quiebra. Evaluar con métricas de performance el modelo de aprendizaje automático para la predicción de la quiebra de empresas.		Hipótesis específicas: El preprocesamiento de la data colectada de empresas optimiza significativamente la predicción del modelo de aprendizaje automático. La evaluación del performance de los algoritmos de aprendizaje automático optimiza significativamente la predicción del modelo de la quiebra de empresas. La correlación de los factores relevantes de las empresas optimiza significativamente el diseño del modelo de aprendizaje automático de la predicción de quiebra. La evaluación con métricas de performance del modelo de aprendizaje automático valida significativamente la predicción de la quiebra de empresas.	Variable Dependiente Quiebra de empresas	

PRODUCCIÓN CIENTÍFICA

Artículo: Revisión Sistemática de la Literatura. Scopus Q4

Título: Prediction model applying machine learning to forecast the bankruptcy of companies. A systematic review of the literature.

Journal of Theoretical and Applied Information Technology

15th August 2024, Vol.102, No. 15

© Little Lion Scientific



ISSN: 1992-8645

www.jatit.org

E-ISSN: 1817-3195

PREDICTION MODEL APPLYING MACHINE LEARNING TO FORECAST THE BANKRUPTCY OF COMPANIES. A SYSTEMATIC REVIEW OF THE LITERATURE.

DINO QUINTEROS-NAVARRO¹, CIRO RODRIGUEZ²

¹Universidad Nacional Mayor de San Marcos, Lima, Peru

²Universidad Nacional Mayor de San Marcos, Lima, Peru

E-mail: ¹dino.quinteros@unmsm.edu.pe, ²crodriguezro@unmsm.edu.pe

ABSTRACT

Prediction models that are aimed at companies allow trends to be identified and generate a much broader picture, making decisions more effective and efficient. In that sense, the research study uses a literature review to identify the state of the art of how predictive models can interact and generate more accurate and reliable results to identify, forecast and eventually reduce the bankruptcy of companies. In addition, the systematic analysis of the contributions was carried out where the main models of supervised learning and other techniques were considered. The authors [25] specify that a systematic literature review is a primary tool for developing an evidence base by identifying, evaluating, and interpreting all available research relevant to a particular research question, thematic area, or phenomenon of interest. Likewise, questions were posed through three (3) stages: identification of parameters, calculation of the ACCP and determination of factors. For the calculation of the weighted precision metric (ACCP), the most significant values of each model were used and the results obtained were analyzed. In this sense, the review must contain the following: method of analysis, theoretical basis, classification of payment yield prediction models and analysis of the topics according to the questions asked.

Keywords: *Bankruptcy, Predictive, Models, Machine Learning, Making Decisions*

Figura 32. Artículo científico publicado. SCOPUS Q4

Fuente. Tomado del trabajo de Quinteros-Navarro y Rodríguez (2024).

<https://www.jatit.org/volumes/hundredtwo15.php>

PRODUCCIÓN CIENTÍFICA

Artículo: Contribución. Scopus Q4

Título: Optimized Deep Learning Model to Predict Business Bankruptcy



Ingénierie des Systèmes d'Information

Vol. 29, No. 5, October, 2024, pp. 1763-1777

Journal homepage: <http://iieta.org/journals/isi>

Optimized Deep Learning Model to Predict Business Bankruptcy

Dino Quinteros-Navarro^{*}, Ciro Rodríguez[†]

Faculty of Systems Engineering, Universidad Nacional Mayor de San Marcos, Lima 15081, Peru

Corresponding Author Email: dino.quinteros@unmsm.edu.pe



Copyright: ©2024 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.290510>

ABSTRACT

Received: 14 March 2024

Revised: 5 August 2024

Accepted: 14 September 2024

Available online: 24 October 2024

Keywords:

deep learning, model, predict, business, bankruptcy

This article studies the ideal and optimized way to predict a company's bankruptcy according to the internal and external environment. Likewise, prognostic factors were identified through machine learning models that are ideal for this type of case. In this context, the choice of the ideal model is the basis for making improvements and updates that allow the optimization of a specific model. For this reason, the optimization of the model is considered a specialized process that allows the precise identification of the factors that lead to the bankruptcy of companies and the identification of the necessary correlations between relevant variables. In this sense, past works on similar contexts were considered for the present study, the analysis method where the methodology is required, and three stages: Data exploration, selection of models, and implementation. Likewise, to determine the results, training was considered to obtain results from the model with optimized characteristics, company bankruptcy factors and correlations, and finally the discussion of results and conclusions was specified.

Figura 33. Artículo científico publicado. SCOPUS Q4

Fuente. Tomado del trabajo de Quinteros-Navarro y Rodríguez (2024).

<https://doi.org/10.18280/isi.290510>

PRODUCCIÓN CIENTÍFICA

INDEXACIÓN SCOPUS. REVISTAS Q4

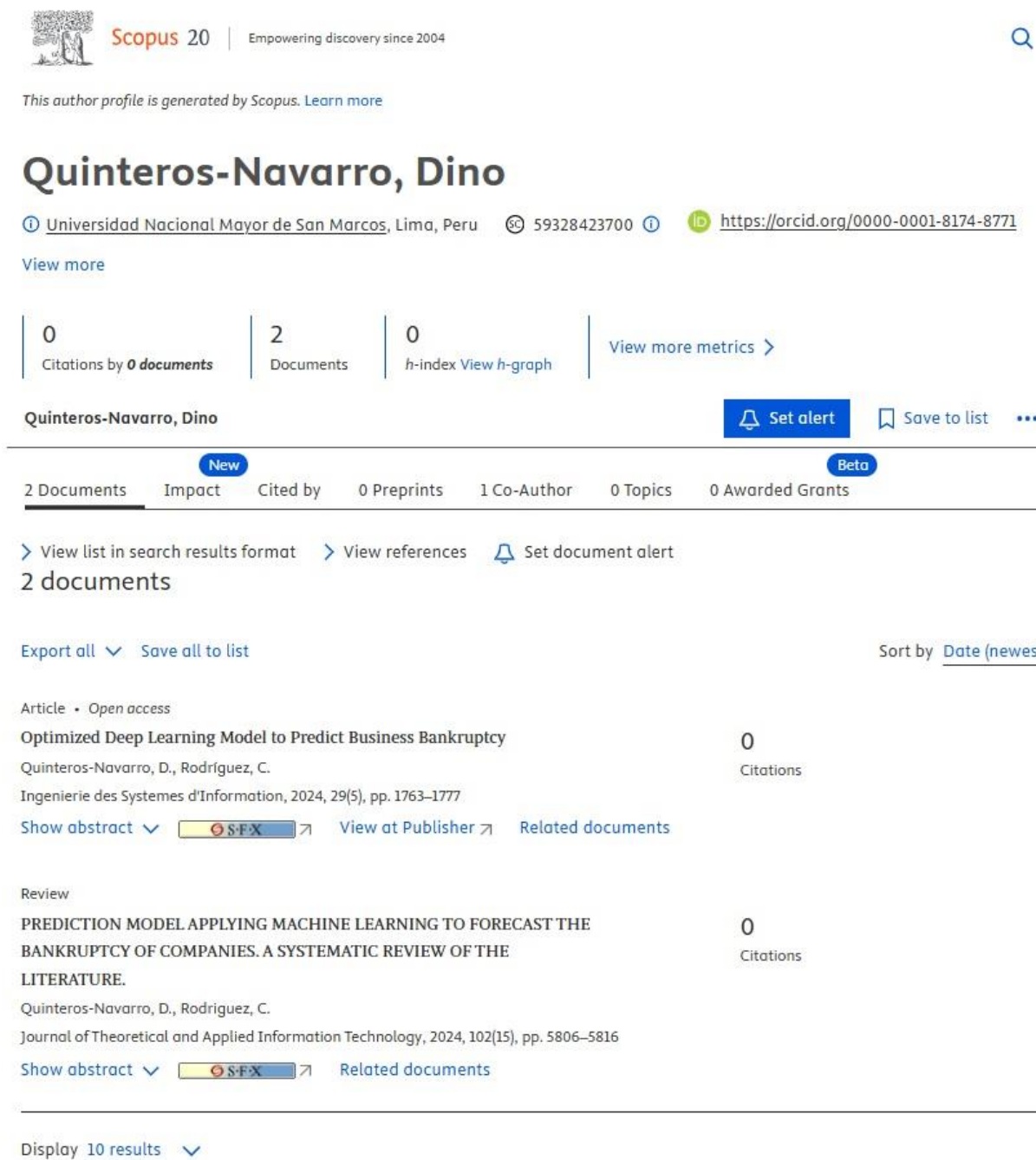


Figura 34. Indexación en SCOPUS. Revistas Q4.

Fuente. Tomado de Elsevier (2024)

<https://www.scopus.com/authid/detail.uri?authorId=59328423700>



Figura 35. Diploma de participación en conferencia internacional (ICWTNS)
Fuente. ICWTNS-2024.

CODIGO EJECUTADO PARA VALIDACIÓN CRUZADA (KFold)

```

✓ [17] X = employee_df.iloc[:, 1:21] # Replace 'data' with 'employee_df'
      y = employee_df['Bankruptcy'] # Replace 'data' with 'employee_df'

✓ [18] from sklearn.model_selection import KFold
      k = 5
      kf = KFold(n_splits=k, shuffle=True, random_state=42)

✓ [19] from sklearn.neural_network import MLPClassifier
      model = MLPClassifier() # Creating an instance of the neural network classifier

✓ [20] from sklearn.model_selection import cross_val_score
      scores = cross_val_score(model, X, y, cv=kf, scoring='r2')
      File "/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_scorer.py", line 388, in _score
        return self._sign * self._score_func(y_true, y_pred, **scoring_kwargs)
      File "/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_regression.py", line 1257, in r2_score
        _check_reg_targets_with_floating_dtype(
      File "/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_regression.py", line 198, in _check_reg_targets_with_floating_dtype
        y_type, y_true, y_pred, multioutput = _check_reg_targets(

[22] import numpy as np

      average_r2 = np.mean(scores)

      print(f"R² Score for each fold: {[round(score, 4) for score in scores]}")
      print(f"Average R² across {k} folds: {average_r2:.2f}")

R² Score for each fold: [nan, nan, nan, nan, nan]
Average R² across 5 folds: nan

# prompt: generar grafica KFold

import matplotlib.pyplot as plt
import numpy as np

X = employee_df.iloc[:, 1:21]
y = employee_df['Bankruptcy']
k = 5
kf = KFold(n_splits=k, shuffle=True, random_state=42)

model = MLPClassifier()
scores = cross_val_score(model, X, y, cv=kf, scoring='accuracy') # Changed scoring to 'accuracy'

average_accuracy = np.mean(scores)

print(f"Accuracy Score for each fold: {[round(score, 4) for score in scores]}")
print(f"Average Accuracy across {k} folds: {average_accuracy:.2f}")

# Generate the plot
plt.figure(figsize=(8, 6))
plt.plot(range(1, k + 1), scores, marker='o', linestyle='--')
plt.title('Accuracy Scores for Each Fold')
plt.xlabel('Fold')
plt.ylabel('Accuracy')
plt.xticks(range(1, k + 1))
plt.grid(True)
plt.show()

```

Figura 36. Código validación cruzada (KFold)

Fuente. Elaboración propia

CODIGO EJECUTADO PARA SHAP – LIME

SHAP

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from google.colab import drive
import shap
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import train_test_split

drive.mount('/content/drive/')
# Debe incluir el enlace completo al archivo csv que contiene su conjunto de datos
Shap = pd.read_csv("/content/drive/My Drive/Data/Shap.csv")
# Assuming X and y are your features and target variable respectively
# Replace 'target_column' with the actual name of your target column
X = Shap.drop('Class', axis=1)
y = Shap['Class']
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Initialize and train the MLPClassifier (replace with your actual model training)
model = MLPClassifier(hidden_layer_sizes=(100,), max_iter=500, random_state=42)
model.fit(X_train, y_train)
# load JS visualization code to notebook
shap.initjs()
# Use KernelExplainer for MLPClassifier
explainer = shap.KernelExplainer(model.predict_proba, X_train)
shap_values = explainer.shap_values(X_test)
# Summary plot
print("Variable Importance Plot - Global Interpretation")
shap.summary_plot(shap_values, X_test)
# Summary plot (bar)
shap.summary_plot(shap_values, X_test, plot_type="bar")
# Waterfall plot for a single prediction
explainer = shap.Explainer(model.predict, X_train)
shap_values = explainer(X_test)
shap.plots.waterfall(shap_values[0])
```

Drive already mounted at /content/drive/; to attempt to forcibly remount, call drive.mount("/content/drive/").

WARNING:shap:Using 303 background data samples could cause slower run times. Consider using shap.si

100% 76/76 [00:36<00:00, 2.64it/s]

Variable Importance Plot - Global Interpretation

Figura 37. Código Shap

Fuente. Elaboración propia

LIME

```
# Explain a single prediction using LIME
# Choose an instance to explain
i = 10 # Example: explain the 10th instance in the test set
# Explain the instance
explanation = explainer.explain_instance(X_test.values[i], model.predict_proba, num_features=10)
# Display the explanation
explanation.show_in_notebook(show_table=True)
# Generate and display the LIME plot
fig = explanation.as_pyplot_figure()
plt.show()
```

Figura 38. Código Lime

Fuente. Elaboración propia

CÓDIGO EJECUTADO PARA EL USO DE IMPUTACIÓN KNN

```

from sklearn.impute import KNNImputer
# Identify columns with missing values
cols_with_missing = employee_df.columns[employee_df.isnull().any()].tolist()

# Check if there are any columns with missing values
if cols_with_missing: # Proceed only if there are columns with missing values
    # Create a KNN imputer
    imputer = KNNImputer(n_neighbors=5) # You can adjust the number of neighbors

    # Fit and transform the imputer on the specified columns
    employee_df[cols_with_missing] = imputer.fit_transform(employee_df[cols_with_missing])

    # Now 'employee_df' has the missing values filled using KNN imputation (if any were present)
    print("Missing values filled using KNN imputation.")
else:
    print("No missing values found in the DataFrame.")

# Print the sum of null values to verify
print(employee_df.isnull().sum())

import matplotlib.pyplot as plt
import seaborn as sns

# Assuming 'employee_df' is your DataFrame and 'cols_with_missing' is defined as before

# Assuming 'employee_df' has been imputed using KNNImputer
# Create a list of columns where we want to examine the distribution
# Replace these with the relevant columns from your DataFrame

for col in ['Months_A', 'Capital', 'Cost_effectiveness', 'Insolvency', 'Satisfaction_E']: # Example columns, replace with yours
    plt.figure(figsize=(12, 7))
    sns.kdeplot(employee_df[employee_df['Bankruptcy'] == 0][col], label='No Bankruptcy', fill=True)
    sns.kdeplot(employee_df[employee_df['Bankruptcy'] == 1][col], label='Bankruptcy', fill=True)
    plt.xlabel(col)
    plt.title(f'Distribution of {col} by Bankruptcy Status')
    plt.legend()
    plt.show()

```

Figura 39. Código imputación KNN

Fuente. Elaboración propia

CODIGO EJECUTADO PARA OBTENER OUTLIERS

```
# Assuming 'employee_df' is your DataFrame

# Identify numerical columns
numerical_cols = employee_df.select_dtypes(include=np.number).columns

# Iterate through numerical columns and detect outliers using IQR
for col in numerical_cols:
    Q1 = employee_df[col].quantile(0.25)
    Q3 = employee_df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Identify outliers
    outliers = employee_df[(employee_df[col] < lower_bound) | (employee_df[col] > upper_bound)]

    # Print or handle outliers (e.g., remove, replace)
    print(f"Outliers in column '{col}':")
    outliers
    # Example: Remove outliers
    # employee_df = employee_df[(employee_df[col] >= lower_bound) & (employee_df[col] <= upper_bound)]

import matplotlib.pyplot as plt
import seaborn as sns

# Assuming 'employee_df' and 'numerical_cols' are defined as in your previous code

for col in numerical_cols:
    plt.figure(figsize=(8, 6)) # Adjust figure size as needed
    sns.boxplot(x=employee_df[col])
    plt.title(f'Box Plot of {col} to detect outliers')
    plt.show()

import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import IsolationForest

# Assuming 'employee_df' and 'numerical_cols' are defined as in your previous code

# Create an IsolationForest model
model = IsolationForest(contamination=0.05) # Adjust contamination as needed

# Fit the model on your numerical data
model.fit(employee_df[numerical_cols])

# Get outlier predictions (-1 for outliers, 1 for inliers)
outlier_predictions = model.predict(employee_df[numerical_cols])

# Add outlier predictions as a column to the DataFrame
employee_df['outlier'] = outlier_predictions

# Visualize outliers using histograms
for col in numerical_cols:
    plt.figure(figsize=(8, 6))
    # Use the 'outlier' column for hue, which is now part of the DataFrame
    sns.histplot(data=employee_df, x=col, hue='outlier', kde=True)
    plt.title(f'Distribution of {col} with Outliers (Isolation Forest)')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.show()
```

Figura 40. Código outliers

Fuente. Elaboración propia

CÓDIGO EJECUTADO PARA LA TÉCNICA DE OPTIMIZACIÓN BAYESIANA (HIPERPARÁMETROS)

```
!pip install scikit-optimize

import skopt
from skopt import gp_minimize
from sklearn.model_selection import cross_val_score
from sklearn.neural_network import MLPClassifier

# Define the objective function to minimize (negative accuracy in this case)
def objective(params):
    # Unpack the parameters
    hidden_layer_sizes, alpha, learning_rate_init = params

    # Create and train the MLPClassifier with the given parameters
    model = MLPClassifier(hidden_layer_sizes=hidden_layer_sizes,
                          alpha=alpha,
                          learning_rate_init=learning_rate_init,
                          random_state=42, # for reproducibility
                          max_iter=1000) # Increase max_iter for better convergence

    # Perform cross-validation
    score = -cross_val_score(model, X_train, y_train, cv=5, scoring='accuracy').mean() #Negative because we minimize

    return score

# Define the search space for the hyperparameters
space = [
    skopt.space.Integer(5, 100, name='hidden_layer_sizes'), # Number of neurons in a hidden layer.
    skopt.space.Real(1e-5, 1, prior='log-uniform', name='alpha'), # L2 penalty (regularization term) parameter.
    skopt.space.Real(1e-5, 1e-1, prior='log-uniform', name='learning_rate_init') # The initial learning rate used.
]

# Perform Bayesian optimization
result = gp_minimize(objective, space, n_calls=50, random_state=0) # Adjust n_calls as needed

# Print the best hyperparameters and the corresponding negative accuracy
print("Best hyperparameters:", result.x)
print("Best negative accuracy:", result.fun)

# Train the final model with the best hyperparameters
best_model = MLPClassifier(hidden_layer_sizes=result.x[0],
                           alpha=result.x[1],
                           learning_rate_init=result.x[2],
                           random_state=42, max_iter=1000)

best_model.fit(X_train, y_train)

y_pred = best_model.predict(X_test)

# Evaluate the model
print(classification_report(y_test, y_pred))
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True)
plt.show()
```

Figura 41. Código optimización bayesiana

Fuente. Elaboración propia

CÓDIGO EJECUTADO PARA LA TÉCNICA DE SELECCIÓN DE CARACTERÍSTICAS (RFE)

```
from sklearn.feature_selection import RFE
import matplotlib.pyplot as plt

# Assuming X and y are your features and target variable
# Replace with your actual data
X = employee_df.drop('Bankruptcy', axis=1) # Features (excluding target)
y = employee_df['Bankruptcy'] # Target variable

# Create a Logistic Regression estimator
estimator = LogisticRegression()

# Create the RFE object (select the number of features you want to keep)
selector = RFE(estimator, n_features_to_select=10, step=1) # Keep 10 features

# Fit the selector to your data
selector = selector.fit(X, y)

# Get the ranking of the features
ranking = selector.ranking_

# Create the plot
plt.figure(figsize=(10, 6))
plt.plot(range(1, len(ranking) + 1), ranking)
plt.xlabel('Feature Index')
plt.ylabel('Feature Ranking')
plt.title('Recursive Feature Elimination (RFE)')
plt.xticks(range(1, len(ranking) + 1))
plt.show()
```

Figura 42. Código selección de características (RFE)

Fuente. Elaboración propia

CÓDIGO EJECUTADO PARA REDUCCIÓN DE DIMENSIONALIDAD (PCA)

```

from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Assuming 'employee_df' is your DataFrame and you want to perform PCA on all variables
# excluding 'Bankruptcy' (or whatever your target variable is named)

# Separate features (X) and target (y)
X = employee_df.drop('Bankruptcy', axis=1)
y = employee_df['Bankruptcy']

# Initialize PCA with the desired number of components
# or set n_components to a value between 0 and 1 to specify the variance explained
pca = PCA(n_components=2) # Example: keep 2 principal components

# Fit PCA on the data and transform the features
X_pca = pca.fit_transform(X)

# Plot the first two principal components
plt.figure(figsize=(8, 6))

# Convert y to numerical values for color coding
y_numeric = y.map({'Yes': 1, 'No': 0}) # Map 'Yes' to 1 and 'No' to 0

plt.scatter(X_pca[:, 0], X_pca[:, 1], c=y_numeric) # Use y_numeric

```

Figura 43. Código reducción de dimensionalidad (PCA)

Fuente. Elaboración propia