UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

TEMA:

Modelo de datos para predecir quiebras de pymes importadoras en Guayaquil: análisis con árboles de decisión y random forest.

AUTOR (ES):

Avila Coello, Milena Jamilet

Mendoza Jumbo, Jair Daniel

Trabajo de titulación previo a la obtención del título de LICENCIADOS EN NEGOCIOS INTERNACIONALES

TUTOR:

Ing. Carrera Buri, Félix Miguel, Mgs.

Guayaquil, Ecuador

12 de febrero del 2026

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de titulación fue realizado en su totalidad por **Avila Coello, Milena Jamilet y Mendoza Jumbo, Jair Daniel**, como requerimiento para la obtención del título de **Licenciado en Negocios Internacionales.**

TUTOR (A)

f. _____

Ing. Carrera Buri, Félix Miguel, Mgs.

DIRECTOR DE LA CARRERA

f. _____

Hurtado Cevallos, Gabriela Elizabeth

Guayaquil, a los 12 del mes de febrero del año 2026

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Yo, **Avila Coello, Milena Jamilet**
Mendoza Jumbo, Jair Daniel

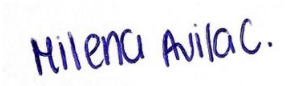
DECLARO QUE:

El Trabajo de Titulación, **(Escriba el tema del trabajo)** previo a la obtención del título de **Modelo de datos para predecir quiebras de pymes importadoras en Guayaquil: análisis con árboles de decisión y random forest**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

Guayaquil, a los 12 del mes de febrero del año 2026

LOS AUTORES



f. _____
Avila Coello, Milena Jamilet

f. _____
Mendoza Jumbo, Jair Daniel

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

AUTORIZACIÓN

Yo, **Avila Coello, Milena Jamilet**
Mendoza Jumbo, Jair Daniel

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación, **Modelo de datos para predecir quiebras de pymes importadoras en Guayaquil: análisis con árboles de decisión y random forest**, cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 12 del mes de febrero del año 2026

LOS AUTORES

f. _____
Avila Coello, Milena Jamilet

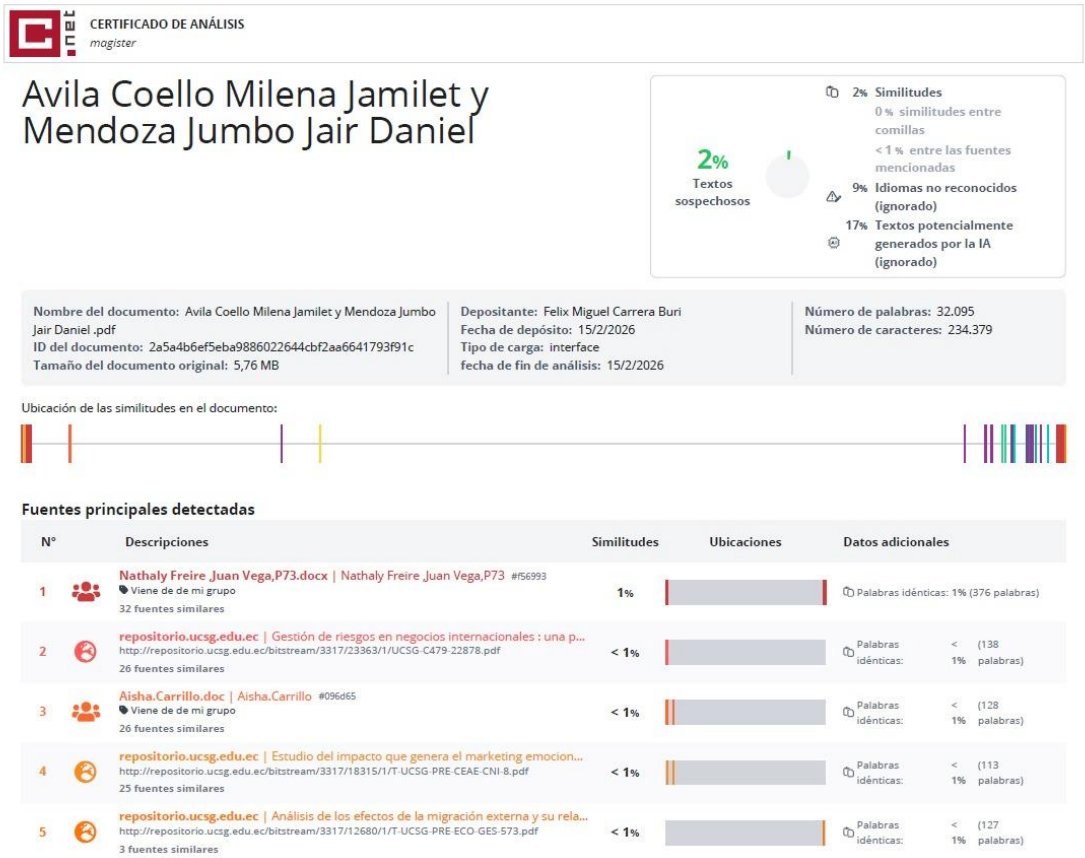
f. _____
Mendoza Jumbo, Jair Daniel

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

Reporte Compilatio



CERTIFICADO DE ANÁLISIS
magister

Avila Coello Milena Jamilet y Mendoza Jumbo Jair Daniel

2% Textos sospechosos

2% Similitudes
0 % similitudes entre comillas
< 1 % entre las fuentes mencionadas
9% Idiomas no reconocidos (ignorado)
17% Textos potencialmente generados por la IA (ignorado)

Nombre del documento: Avila Coello Milena Jamilet y Mendoza Jumbo Jair Daniel .pdf
ID del documento: 2a5a4b6ef5eba9886022644cbf2aa6641793f91c
Tamaño del documento original: 5,76 MB

Depositante: Felix Miguel Carrera Buri
Fecha de depósito: 15/2/2026
Tipo de carga: interface
fecha de fin de análisis: 15/2/2026

Número de palabras: 32.095
Número de caracteres: 234.379

Ubicación de las similitudes en el documento:

Fuentes principales detectadas

N°	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	Nathaly Freire ,Juan Vega,P73.docx | Nathaly Freire ,Juan Vega,P73 #f56993 Viene de de mi grupo 32 fuentes similares	1%		Palabras idénticas: 1% (376 palabras)
2	repositorio.ucsg.edu.ec | Gestión de riesgos en negocios internacionales : una p... http://repositorio.ucsg.edu.ec/bitstream/3317/23363/1/UCSG-C479-22878.pdf 26 fuentes similares	< 1%		Palabras idénticas: < 1% (138 palabras)
3	Aisha.Carrillo.doc | Aisha.Carrillo #096d65 Viene de de mi grupo 26 fuentes similares	< 1%		Palabras idénticas: < 1% (128 palabras)
4	repositorio.ucsg.edu.ec | Estudio del impacto que genera el marketing emocion... http://repositorio.ucsg.edu.ec/bitstream/3317/18315/1/T-UCSG-PRE-CEAE-CNI-8.pdf 25 fuentes similares	< 1%		Palabras idénticas: < 1% (113 palabras)
5	repositorio.ucsg.edu.ec | Análisis de los efectos de la migración externa y su rela... http://repositorio.ucsg.edu.ec/bitstream/3317/12680/1/T-UCSG-PRE-ECO-GES-573.pdf 3 fuentes similares	< 1%		Palabras idénticas: < 1% (127 palabras)

TUTOR (A)

f. _____

Ing. Carrera Buri, Félix Miguel, Mgs.

AGRADECIMIENTO DE MILENA AVILA

A Dios, por ser mi guía en cada etapa de este proceso y por darme la fortaleza que necesitaba cuando pensé en rendirme.

A mi papá, Hicler Avila, por su apoyo y amor incondicional, por su sacrificio año a año para que yo obtuviera la mejor educación y porque muchas de las oportunidades que tuve fueron gracias a su esfuerzo.

A mi mamá y mejor amiga, Fernanda Coello, por inculcarme desde muy pequeña lo valioso e importante que es obtener mi título profesional, por sus palabras de apoyo cuando sentía que no podía, por la ilusión y amor que sostuvo este sueño, y por creer en mí incluso cuando yo dudaba.

A mi novio, Luis Alejandro Correa, por acompañarme en este proceso, por ayudarme a superar cada obstáculo, por impulsarme a continuar, por sostenerme en los momentos más difíciles de esta etapa y por celebrar cada logro mío como si fuera suyo.

A mi hermano, Rubén Avila, por su paciencia y apoyo llevándome y recogiéndome de la universidad, y por estar para mí a pesar de todo.

Al Ing. Félix Carrera, por haber sido un excelente tutor, por su paciencia, su guía y su disposición por ayudarnos durante este proceso. Gracias por compartir sus conocimientos y ser una parte fundamental para la culminación de este trabajo.

Finalmente, a todas las personas que fueron parte de mi etapa universitaria, los llevaré conmigo siempre.

DEDICATORIA DE MILENA AVILA

Dedico este trabajo de titulación a mis padres, Hicler Avila y Fernanda Coello, quienes se esforzaron día a día para darme lo mejor, porque sin su apoyo nada de esto hubiera sido posible, este logro también es de ellos.

AGRADECIMIENTO DE JAIR MENDOZA

Este logro es el esfuerzo y el apoyo de todas las personas que han sido parte de mi vida, personas que me impulsaron a seguir adelante y a no rendirme, en donde compartí momentos inolvidables juntos a ellos, risas, tristezas, alegrías muchos sentimientos que se podría expresar de una manera y es el amor.

A mis Abuelos Sambuco y la yiyi quienes me vieron crecer y me han ayudado en todo, sib personas que me han apoyado siempre y me enseñaron a nunca rendirme y agradezco todo ese amor y apoyo incondicional que han dado.

A mi Abuela Tata, agradezco de todo corazón el siempre estar velando por mi salud, una persona que me crio de pequeño como a un hijo, que me dio amor y siempre estuvo ahí para mí, te amo demasiado y sé que todo lo que tengo hasta ahora es gracias a ti a tu ayuda silenciosa, Por eso sé que tu amor y toda la que me has dado es lo más lindo que me ha llegado pasar.

A mi Abuelo Baba, una persona a la que ame demasiado, una persona que siempre estuvo conmigo, quien me acompaño a cualquier lado para comprar un helado o cosas de la universidad, sé que me demostraste demasiado amor y pude contar contigo para todo y cuando salimos con la carreta a vender comida fue lo más lindo que llegue a experimentar porque se supe de donde viene el dinero y lo mucho que cuesta, desde ese día que te fuiste, fue un evento para mi muy doloroso porque uno de mis pilares de mi vida se derrumbó y no supe que hacer, pero como tú me enseñaste a como levantarme y a seguir adelante cuando todo está en tu contra. Por eso te dedico mi tesis, porque no solo es mi fuerzo es el esfuerzo de todos y más el tuyo porque sé que desde el cielo siempre está velando por mi bienestar y seguridad, por eso quiero que me veas triunfar en la vida.

A mis tíos, Marco y Viviana que siempre estuvieron ahí conmigo para lo que sea, me han ayudado demasiado y agradezco ese amor lindo que me dieron y siempre apoyándome en todo y sé que siempre han velado por mi bienestar y mi salud, y a mi primo Farid, eres una personita muy adorable y

tierna, te quiero demasiado y gracias por hacerme reír con tus ocurrencias e ingenuidad

A mi Tía Mamo agradezco ese amor incondicional que me dio desde la niñez y hasta la actualidad quien siempre ha estado ahí para mí y quien me enseño que todo es posible si vamos de la mano de nuestro señor Jehová, un amor materno que siempre voy a atesorar, siempre me ayudo a que nunca me rinda por eso sé que sus palabras y consejos siempre los voy a escuchar y a ponerlos en práctica.

A mi padre que siempre me demostraron que lo más importante que uno tiene en esta vida es la familia, que siempre ellos te van a apoyar en todo, mi padre un hombre que me enseño lo que es la vida, mi ejemplo a seguir por su esfuerzo y dedicación que da, gracias a ti he logrado un millón de cosas, sé que te esforzaste mucho para que yo pudiera llegar hasta aquí, gracias de corazón por soportar todas mis inmadures que demuestro y aun así me has apoyas. Gracias a todos los consejos que me has dado gracias a eso sé que puedo seguir adelante sin derrumbarme.

A mi mamá una mujer que me demostró lo que es el amor de una madre, gracias por todo ese cariño y amor que me diste agradezco el tiempo y dedicación que me diste, ayudándome en todo momento, eres uno de los pilares de mi vida porque sé que si no estas, no sabría como seguir adelante. Por eso agradezco de corazón, y los sacrificios que mi papa y tu han hecho por mí y por los momentos que hemos pasado, pero siempre juntos y saliendo adelante.

A mi hermana, quien con sus locuras siempre ha estado ahí para mí, una personita que con sus ocurrencias me hace reír y que siempre me esperaba hasta tarde cuando llegaba de una fiesta gracias por todo esa dedicación, apoyo y amor que me has dado, sé que soy tu ejemplo a seguir, te deseo que siempre tengas lo mejor del mundo y con la bendición de Dios sé que todo es posible en esta vida.

A mis amigos Ider Rivadeneira, David Luzardo, Carlos Salinas, William Agurto, Fiorella Valle, Giannella Murillo y Genesis Quinteros, para mi

ustedes son muy apreciados porque siempre estuvieron en todo conmigo, cuando me derrumbaba por amor o por pequeñas cosas siempre estuvieron ahí conmigo, me aconsejaron, me ayudaron cuando más lo necesitaba por eso los aprecio demasiado. Por eso quiero que sepan que siempre van a contar conmigo para lo que sea y sé que así mismo puedo contar con sus ayudas.

A mi mejor amiga Dennise Lindao, una persona que siempre estuvo conmigo desde el inicio una persona que nunca se fue y siempre estuvo ahí, te quiero demasiado y gracias por todo tu tiempo, amor y apoyo que me has dado, eres uno de mis pilares de mi vida y sé que contigo puedo contar para todo. Gracias de corazón mi incondicional.

Agradezco a Angie Gutiérrez, gracias a ese amor que me has dado, ese amor que siempre anhele y el cariño que me has demostrado demasiado, gracias a ti por ayudarme en todo, amarme en todos los momentos. Tu amor sincero y dedicación me han llevado a seguir adelante en todo a nunca rendirme a demostrar que puedo con todo que no existe peros que todo se puede en esta vida. Te amo con todo mi corazón mi princesa.

DEDICATORIA DE JAIR MENDOZA

A Jehová, por ser mi guía espiritual y fortaleza en cada momento de este proceso que me ha ayudado a seguir adelante en todo. Porque las dudas que tenía y los problemas que surgieron el me los resolvió, en mis estados de ánimo más bajos hallé fuerzas y en mis miedos, esperanza. Sin sus oraciones, esto no habría sido posible.

A mis padres, quienes con amor infinito y sacrificios me han enseñado que todo se puede en esta vida con disciplina y fe y que nunca olvide que la familia siempre va a ir primero. Este logro también es suyo, gracias al apoyo incondicional que siempre me brindaron.

A mi abuelo baba que con su dedicación y sabiduría han sido faros de luz en mi camino. A él le dedico mi Tesis y gracias por ser mi motor en todo y siempre le demostrare que si puedo en esta vida y que nada me detendrá

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

TRIBUNAL DE SUSTENTACIÓN

f. _____

(NOMBRES Y APELLIDOS)

TUTOR

f. _____

(NOMBRES Y APELLIDOS)

DECANO O DIRECTOR DE CARRERA

f. _____

(NOMBRES Y APELLIDOS)

COORDINADOR DEL ÁREA O DOCENTE DE LA CARRERA

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Esta investigación evalúa el riesgo de quiebra operativo vinculado a la presión financiera en una compañía importadora de Guayaquil (FEGOCAR) por medio de aprendizaje automático, utilizando unas herramientas de alerta para el control interno. Se entrenó un clasificador supervisado con 4531 transacciones entre 2021 y 2025 empleando variables económicas y operativas, por ejemplo precio transporte seguro cantidad peso neto país de origen el tipo de mercancía modelo y además proveedor. Se llevó a cabo la limpieza de datos, la conversión del formato monetario a un estándar, la codificación one-hot de las variables categóricas y el desarrollo de métricas derivadas: precio_unitario, peso_unitario y costo_logistico_ratio, siendo que para clasificar operaciones como de alto o bajo riesgo, se entrenaron un Random Forest y un Árbol de Decisión con equilibrio entre las clases. El KPI de riesgo operativo se ubicó en 48,78% (2.210 transacciones en alto riesgo). Forest logró AUC 0,824 y exactitud 0,827; el Árbol de Decisión logró AUC 0,817 y exactitud 0,819. La importancia de variables dio como predictores de mayor contribución a peso_unitario, costo_logistico_ratio y precio_unitario. Se generaron rankings por proveedor: tasa de riesgo y gasto total, así como reglas legibles del árbol y una calculadora para estimar el riesgo en nuevas transacciones, útil para auditoría focalizada, validación de costos logísticos y control de transacciones inusuales.

Palabras clave: *Random Forest, Árbol de Decisiones, predecir quiebras, Phyton, economía*

ABSTRACT

This research assesses the risk of operational bankruptcy linked to financial pressure in an import company in Guayaquil (FEGOCAR) through machine learning, using an alert tool for internal control. A supervised classifier was trained with 4,531 transactions between 2021 and 2025 using economic and operational variables such as price, transport, insurance, quantity, net weight, country of origin, type of goods, model, and supplier. Data cleaning, conversion of the monetary format to a standard, one-hot encoding of categorical variables, and the development of derived metrics (unit_price, unit_weight, and logistics_cost_ratio) were carried out. To classify operations as high or low risk, a Random Forest and a Decision Tree with class balance were trained. The operational risk KPI stood at 48.78% (2,210 high-risk transactions). Forest achieved an AUC of 0.824 and an accuracy of 0.827; the Decision Tree achieved an AUC of 0.817 and an accuracy of 0.819. The importance of variables yielded the following predictors with the greatest contribution: unit_weight, logistics_cost_ratio, and unit_price. Rankings were generated by supplier: risk rate and total expenditure, as well as readable tree rules and a calculator to estimate the risk in new transactions, useful for targeted auditing, logistics cost validation, and control of unusual transactions.

Keywords: *Random Forest, Decision Tree, predicting bankruptcies, Python, economics*

INTRODUCCIÓN

Parte productiva de la ciudad de Guayaquil se sostiene en gran medida por pequeñas y medianas empresas a las que se denominan (PYMES), mismas que cumplen un rol importante en actividades de comercio exterior, principalmente en la importación de insumos y bienes producidos (Valente Galán, 2024) Cada una de estas unidades empresariales, se han convertido en pilar fundamental para la generación de empleo, incremento del dinamismo comercial y cadena de suministro local; no obstante, estas organizaciones sufren cambios constantes principalmente en el entorno macroeconómico, como limitaciones a crédito, diversificaciones e la demanda y problemas de liquidez, las cuales pueden llevar a estos establecimientos a procesos de insolvencia o quiebra.

En la actualidad el mercado global es altamente competitivo, y se encuentra marcado por cambios en los precios internacionales costos logísticos indeterminados, rigidez en la cadena de suministro y variaciones en las políticas de comercio exterior, situaciones en las cuales las PYMES importadoras enfrentan estos tipos de escenarios, donde las decisiones comerciales deben ser tomadas con información veraz y oportuna. Estudios desarrollados por la CEPAL (2023) se reconoce que las PYMES latinoamericanas presentan menor capacidad de resiliencia ante crisis económicas y menor acceso a herramientas tecnológicas para gestionar riesgos financieros y operativos, comparadas con empresas de mayor tamaño.

Es así como, en ciertos países de América Latina como Colombia, México y Chile se han realizado investigaciones basadas en la aplicación de modelos de aprendizaje automático para prevenir quiebras empresariales cuyos niveles de precisión han sido superiores al 85% (Terreno et al., 2024). Se estima que cada investigación ejecutada contribuya a la detección a tiempo de las posibles causas que genera la falta de liquidez en las pequeñas empresas, lo que garantizara su continuidad operacional en el mercado comercial.

Dentro del contexto ecuatoriano Andino et al. (2022) evidencia que las PYMES, constantemente enfrentan limitaciones de carácter financiero,

inconsistencias en la gestión económica y mucha presión causada por la crisis monetaria, las cuales han afectado el progreso y la sostenibilidad. Las PYMES en el ecuador tienen un rol determinante dentro de la economía. De acuerdo con datos presentados por el Ministerio de Producción (2025), el 90% del movimiento empresarial del país es generado por las pequeñas y medianas empresas, generando cerca del 60% de empleo para los ciudadanos.

En las principales ciudades del Ecuador una de ellas Guayaquil, el porcentaje de las PYMEs, es realmente representativo, debido al movimiento económico diario por la actividad portuaria y comercial, la misma que se convierte en brecha para el ingreso de mercancía derivadas de América, Asia y Europa. Sin embargo, pese al nivel de importancia que representan las PYMEs a la económica local, estas organizaciones son altamente susceptibles a diversos eventos externos, principalmente aquellos que dependen de actividades de importación (Andino et al., 2022).

En la actualidad en el Ecuador principalmente en Guayaquil, no existen investigaciones sobre el uso de Modelo de datos para predecir quiebras en las PYMEs, las mismas que permitan evaluar los riesgos financieros. Fernández et al. (2024) señalan que uno de los principales factores para que no se desarrollen este tipo de modelos es la escasa cultura en el uso de datos de la gestión gerencial, agregado a esto el temor que tienen las empresas a la hora de compartir información financiera. De esta manera se puede identificar que no existe dicho modelo diseñado utilizando algoritmos como Random Forest y arboles de decisión para la realidad económica de las PYMEs de esta ciudad.

Según Osinaga Flores (2024) las pequeñas y medianas empresas enfrentan presiones principalmente en los costos logísticos y aumento en el despacho aduanero, estas dificultades se evidencian al momento de acceder a financiamiento a tasas competitivas. Esto provoca en reiteradas ocasiones que el 55% de las PYMEs cierren sus operaciones sin haber detectado con anticipación señales de alerta financiera

El proceso de insolvencia no ocurre de la noche a la mañana, sino que se va desarrollando de forma progresiva mediante señales que pueden identificarse en los estados financieros. Siendo una de las posibilidades más relevantes de detección a tiempo los modelos predictivos. Los cuales

representan una ventaja competitiva para las PYMEs importadoras, permitiendo tomar decisiones acertadas antes de llegar a un punto de no retorno.

Entre las técnicas que han dado mayores resultados están los árboles de decisión y algoritmos como Random Forest, los cuales tienen la capacidad de manejar múltiples variables y encontrar patrones que muchos modelos estadísticos no han logrado identificar. Según el autor Espinoza Zúñiga (2020) estos algoritmos no solo predicen si una empresa está atravesando una situación de quiebra, sino que también muestra cuales son las variables que influyen en esta predicción, lo que permite obtener un panorama más ilustrativo y con conocimiento útil para la toma de decisiones.

La aplicación de un modelo de datos predictivos permitirá reconocer patrones que están vinculados a los riesgos que pueden tener las PYMEs, asociados a las variables financieras de liquidez, rentabilidad, endeudamiento, gastos operativos y rotación de inventarios, de la misma forma las variables operativas que se relacionan con frecuencia de importación, dependencia de proveedores internacionales y comportamiento del ciclo de caja. Para hallar dichas variables es necesario los modelos de árbol de decisión, mismos que permiten generar reglas claras de interpretación, mientras que Random Forest mejora la precisión reduciendo el riesgo de sobreajuste, el cual es un problema presente en modelos de predicción financiera. Al aplicar dichos mecanismos se obtendrá como resultado una herramienta capaz de clasificar a las empresas en dos categorías: riesgo de quiebra o no riesgo, permitiendo actuar de forma preventiva y anticipada.

Con esta aplicación se establece una relevancia a la hora de anticipar la quiebra empresarial, por medio de la utilización de los modelos predictivos. Hoy en día las empresas, operan en un ambiente dinámico caracterizado por la digitalización acelerada, esta acción ha incrementado la competencia internacional facilitando el acceso a grandes cantidades de información. En la actualidad la capacidad de analizar datos en tiempo real y predecir el comportamiento a futuro ya no es una ventaja sino requisito paras que las pequeñas y medianas empresa sobrevivan.

Según un informe presentado por el Banco Interamericano de Desarrollo (BID, 2024), se indica que las empresas incorporan modelos

predictivos en la toma de decisiones, y que, esta acción incrementa hasta un 35% su probabilidad de sostenibilidad y permanencia en el mercado. Sin embargo, a pesar de estas evidencias, gran parte de las PYMEs ecuatorianas aún toma decisiones basadas en experiencia, intuición o análisis manual de información financiera. Esta situación abre camino hacia la disponibilidad de datos y uso efectivo de los mismos, abriendo una oportunidad para el uso de algoritmos como árboles de decisión y Random Forest, capaces de transformar datos dispersos en conocimiento estratégico.

La aplicación de estos modelos de predicción de quiebra tiene gran importancia principalmente cuando se analiza el comportamiento del sector importador. Adan et al. (2022) evidencia que el 62% de las PYMEs importadoras presentan inestabilidad en sus ciclos de caja, causadas por los retrasos a la hora de nacionalizar la mercancía, irresoluciones de costos logísticos y variaciones inesperadas en los tiempos de entrega. Esto ha provocado distintas tensiones financieras que perjudican a la e incrementan el riesgo de insolvencia.

Los modelos de predicción de quiebra no se sustentan en adivinar el futuro, sino en reconocer los patrones financieros e interpretar los datos para anticiparse a escenarios previos. A diferencia de los métodos tradicionales, los modelos Random Forest pueden actualizarse a medida que se alimentan con nuevos datos financieros y operativos. Además, permiten medir la importancia de cada variable involucrada en el riesgo de quiebra, lo que contribuye a identificar prácticas empresariales que deben corregirse, la gran ventaja del aprendizaje automático radica en su capacidad de adaptarse a entornos cambiantes y en que sus resultados son cuantificables y verificables.

De esta manera este estudio identificó que la mayoría de las PYMEs importadoras en Guayaquil toman decisiones financieras basadas en intuición o métodos tradicionales de análisis contable, los cuales permiten observar el problema solo cuando ya es evidente y difícil de revertir. A lo largo del proceso de revisión documental y exploración inicial de datos, se evidenció que estas empresas rara vez emplean modelos predictivos que anticipen el riesgo de quiebra con base en patrones históricos. Debido a esta realidad, se propone la creación de un denominado modelo de aprendizaje automatizado, el mismo en el que se apliquen algoritmos como árboles de decisión y Random Forest,

ya que su funcionabilidad y capacidad permite la identificación de posibles variables económicas, que puedan llegar a predecir riesgos financieros comprensibles para las empresas,

La utilización de este sistema proporcionará indicadores financieros que ocasionan los riesgos de insolvencia económica y peligro de quiebras, por lo cual este instrumento se constituye como un valioso aporte para las gestiones organizacionales. De tal forma, este estudio no solo plantea la predicción de un posible escenario de quiebra, sino también la comprensión de sus causas, lo que favorece la continuidad empresarial en mercados altamente competitivos.

PROBLEMÁTICA

En la ciudad de Guayaquil-Ecuador, las pequeñas y medianas empresas que se dedican a la importación representan un sector muy clave para la economía local. Sin embargo, muchas empresas también pueden representar altos niveles de vulnerabilidad financiera debido a que el mundo está en constante innovación y todo esto genera algún tipo de cambio como las variaciones en los costos logísticos, dependencia de proveedores internacionales y esto ocasiona un problema en donde es muy limitado y el riesgo incrementa. Estas condiciones han generado un incremento sostenido en los índices de las quiebras empresariales, afectó al mercado y el poder generar empleos.

Tradicionalmente, el análisis financiero de las PYMES se basa en métodos descriptivos o en la revisión de los indicadores contables, los cuales con estos datos no se pueden identificar de manera oportuna ya que los datos obtenidos no son limitados.

De acuerdo con el comex, en la actualidad surge utilizar el machine learning, que permite procesar grandes volúmenes de datos y descubrir relaciones entre las variables logísticas y externas que puedan perjudicar a que se lleve una quiebra (Comex, 2022).

De esta manera, se han consolidado como unas herramientas más potentes para extraer conocimientos a partir de grandes volúmenes de información para realizar predicciones precisas sobre comportamientos

futuros. En el ámbito empresarial, estas metodologías permiten anticipar riesgos, optimizar procesos y mejorar las tomas de decisiones.

El machine learning, al ser un concepto de inteligencia artificial, ofrece algunas alternativas modernas y potentes para enfrentar esta problemática. A través de sus algoritmos por lo que las empresas poseen una gran base de datos financiera, comercial y logística (Kelleher & Tierney, 2018), gracias a esta herramienta permite encontrar patrones para predecir comportamientos futuros con un alto grado de precisión sobre sus importaciones. Sin embargo, a pesar de tener un gran potencial, la adopción de este método sigue siendo muy baja debido a varios factores que se deben a que las personas no tienen tantos conocimientos sobre la inteligencia artificial y esto perjudica mucho a su personal ya que no tienen el conocimiento suficiente, ausencia de infraestructura tecnológica y percepción errónea de que su implementación sólo es viable en las empresas que tengan un gran volumen de datos.

La brecha que existe entre las pymes de las importadoras en Guayaquil y las tendencias globales en analítica de datos es muy evidente. En otros países el uso del machine learning se ha convertido en un componente esencial para la competitividad, en el contexto local las empresas guayaquileñas continúan dependiendo de hojas de cálculos, reportes contables y sistemas administrativos básico que no permiten un mayor enfoque en el análisis predictivo ni la autorización inteligente te la toma de decisiones. Este problema que es debido a la falta de conocimientos y capacitaciones no solo limita a la eficiencia operativa, sino que también perjudica a las adaptaciones que haya en el mundo antes las crisis financiera o fluctuaciones del mercado internacional.

A pesar de los avances internacionales en el uso de la inteligencia artificial aplicada en la predicción de insolvencia, en el contexto ecuatoriano y más específico en la ciudad de Guayaquil, la incorporación de estas herramientas sigue siendo muy limitada. Las Pymes que se dedican a la importación carecen de modelos de análisis predictivo que les pueda anticipar su riesgo financiero y adoptar medidas preventivas (Villamar & Cedeño, 2021). En esta mayoría de casos, los empresarios no disponen de una información muy concreta ni de modelos automatizados que integren datos históricos, indicadores financieros y variables macroeconómicas para realizar un

diagnóstico más sofisticado. De igual forma, estas instituciones financieras otorgan créditos a estas empresas no siempre cuentan con herramientas que se utilicen en el machine learning que faciliten a la predicción de riesgo que reduzcan las importaciones del sector.

La escasez de modelos predictivos de quiebra establecidos en la aplicación de machine learning forja una brecha significativa entre el potencial de las tecnologías existentes y la realidad operativa de las pymes importadoras del mercado local. Esto obstaculiza los beneficios del análisis de datos, como la optimización de decisiones, prevención de crisis y mejora de la sostenibilidad empresarial, Villamar, M., & Cedeño, K. (2021).

Además, la falta de investigación local en torno de la aplicación de algoritmos como los árboles de decisión y random forest limita la posibilidad de construir conocimiento contextualizado, ajustado a las características propias que difiere notablemente de otros entornos internacionales por factores económicos, culturales y financieros.

Estos modelos son capaces de procesar grandes volúmenes de información, analizar simultáneamente múltiples variables y generar varios modelos predictivos que dependen de sus comportamientos pasados para poder identificar un punto de quiebre que haya tenido la empresa o si se ha mantenido estable. Estos algoritmos no solo clasifican el nivel de riesgo, si no también ofrecen interpretaciones de manera visual y jerárquica de las variables más influyentes en el resultado, ayudando a la compresión por parte de los empresarios y analistas financieros locales.

Este problema se agrava debido a que la detección de señales tempranas de crisis financieras no suele formar parte de la cultura empresarial de las pymes locales. La mayoría de las decisiones se basan solo en la intuición o en análisis retrospectivos, sin una base de datos ya que con esto se puede predecir y permitir anticipar posibles escenarios adversos. En consecuencia, muchas empresas locales suelen reaccionar de manera tardía y eso perjudica a que tengan pérdidas que son irreversibles.

Finalmente, se hace necesario desarrollar modelos de predicciones basados en algoritmos de machine learning que se adapten a la realidad empresarial guayaquileña, integrando una información más rentable y operativa dando a que el estudio del mercado seas más fácil de que pueda

generar predicciones útiles que fortalezcan la gestión y sostenibilidad de las empresas locales, “Modelo de predicción de quiebra en empresas de comercio en Ecuador” (Molina et al., 2023).

La utilización de estas herramientas no solo representaría un avance en la digitalización del sector productivo, sino también una oportunidad para posicionar a Guayaquil como un referente en la aplicación de la inteligencia artificial al ámbito empresarial.

Justificación

En la actualidad el entorno económico, está marcado por la globalización comercial y el acelerado avance tecnológico, mismo que ha generado nuevos desafíos para la sostenibilidad de las (PYMES) importadoras que forman parte de la ciudad de Guayaquil. Estas organizaciones mantienen una intervención significativa en la actividad empresarial del país, aportando dinamismo al flujo comercial del puerto y generando empleo directo e indirecto (Yance et al., 2021).

Según información proporcionada por la (ONU, 2024) la estructura financiera de este tipo de empresas suele ser sensible a variaciones externas, principalmente a los cambios asociados al comercio internacional: tiempos de despacho, fluctuación en los costos logísticos, variación en el tipo de cambio y exigencias regulatorias en procesos aduaneros. Estas condiciones hacen que cualquier alteración influya directamente en el flujo de caja, lo que incrementa el riesgo de insolvencia y, en casos más críticos, terminan en procesos de quiebra empresarial

Aunque las PYMES representan una parte significativa de la economía local, se ha podido identificar que la mayoría de ellas no dispone de mecanismos preventivos que permitan anticipar señales de alerta financiera. A menudo, las decisiones estratégicas se toman únicamente en función de la experiencia o de análisis contables estáticos, los cuales ofrecen una visión retrospectiva de los hechos y no permiten detectar eventos que están por desarrollarse (Guamán y Goya, 2025).

Esta implementación en la gestión comercial se considera como reactiva en lugar de ser preventiva, lo que limita la capacidad de respuesta ante situaciones contrarias. Cuando las consecuencias financieras son evidentes, probablemente ya exista un deterioro en la liquidez, la forma de

pago y las relaciones entre los proveedores internacionales, lo cual se refleja en la poca capacidad de supervivencia que muestra la empresa.

De acuerdo con esta realidad, es que nace la necesidad de aplicar herramientas analíticas que permitan determinar a tiempo escenarios adversos y detectar esquemas de riesgos, antes de que la situación se torne complejo e impida la no prevención. La utilización de modelos predictivos, a través del aprendizaje automatizado se considera como una herramienta poderosa que han reflejado efectividad principalmente en economías desarrolladas. Dichos modelos analizan sistemáticamente diversas variables y las relaciona con los modelos tradicionales, de modo que se pueda determinar si son o no efectivas en función de los modelos actuales. Los modelos más destacados son los árboles de decisión y el algoritmo Random Forest, mismos que por su capacidad pueden utilizar las empresas para medir la probabilidad de escenarios de quiebra.

El aporte de esta investigación radica en que la predicción de riesgo no es limitada solo a insolvencias o falta de liquidez, sino que también identifica cuales son los indicadores económicos que influyen con mayor peso en el declive financiero. Las variables o indicadores que se pueden identificar como resultado de este análisis son: liquidez, endeudamiento, rentabilidad, rotación de inventario o dependencia de proveedores externos; estos causan un impacto significativo en la toma de decisiones preventivas antes de decidir de forma definitiva. La utilidad del modelo no radica únicamente en su capacidad para producir un resultado final, sino en su contribución al análisis interpretativo, brindando información clara que puede ser utilizada por gerentes, contadores y responsables de logística o finanzas.

Es así como, desde una ámbito académico, este estudio representa una contribución significativa debido a la carencia de investigaciones sobre este tema en el ámbito local. Pese a que otros países de la región ya se han implementado modelos de aprendizaje automático para analizar quiebras, en el Ecuador el uso de modelos predictivos aún es limitado. En Guayaquil, el sector importador aun es débil, debido a la falta de información y conocimiento sobre los modelos de prevención, razón por la cual los empresarios han empezado a actuar con gran preocupación, lo que ha provocado un retroceso

en el uso de herramientas tecnológicas, las cuales son actualmente en otros países esenciales para el funcionamiento operativo de la economía local.

El desarrollo de un modelo predictivo principalmente aplicado a las PYMES importadoras es indispensable para obtener beneficios prácticos, debido al funcionamiento de las herramientas y a la información que proporcionan de forma estratégica, las mismas que permiten mejorar los procesos en la toma de decisiones. Cambiando de una forma completa la reacción negativa que pueden generar los planes de prevención con datos reales.

Cuando una organización pierde su capital, no solo se afecta económicamente el propietario sino todo el equipo de trabajo, proveedores, y sector económico. Esta quiebra genera pérdida de empleo, desaceleración comercial y disminución económica. Si lo relacionamos a la ciudad de Guayaquil, donde por su movimiento económico se considera como una de las localidades más productivas del Ecuador, es importante mantener un buen sistema de análisis de la realidad de las PYMES, ya que ella son una clave para las oportunidades laborales. Es por esta razón, que disponer de una herramienta predictiva disminuye el porcentaje de quiebras y repercute directa en la estabilidad económica de la ciudad.

Este proyecto investigativo se evidencia desde la perspectiva de innovación. Ya que integra técnicas como árboles de decisión y Random Forest, mismas que aportan hacia un enfoque moderno al análisis financiero, dejando de lado la dependencia exclusiva de modelos contables tradicionales. Ambos algoritmos permiten trabajar con grandes volúmenes de datos y generar resultados con altos niveles de precisión. Además, Random Forest reduce el riesgo de sobreajuste al combinar múltiples árboles, lo que ofrece una mayor fiabilidad en los resultados. Esto convierte al modelo en una herramienta factible, capaz de actualizarse con nueva información y adaptarse al comportamiento cambiante del mercado.

De esta forma esta investigación radica en la posibilidad de convertir datos dispersos en conocimiento estratégico. La información financiera deja de ser un simple registro histórico para convertirse en un insumo que guía decisiones futuras. Con este modelo, las empresas no solo conocerán su condición actual, sino que podrán visualizar posibles escenarios de riesgo,

establecer estrategias de acción y fortalecer su sostenibilidad en un entorno empresarial cada vez más competitivo.

ALCANCE

El presente estudio tiene como propósito fundamental el desarrollo de un modelo de datos para la predicción de quiebras de importadoras en Guayaquil. A partir del uso de algoritmos de machine learning como árboles de decisiones y random forest. En donde se busca construir un sistema para poder analizar el patrón financiero, logísticos y operativos en donde se llegue a conocer si hay un desbalance en las finanzas de la empresa y esto puede perjudicar a que se cierre o tengan déficit de dinero al momento de importar.

Este modelo ayuda a que sea una herramienta de apoyo para las tomas de decisiones estratégicas tanto por parte de los administrativos de las empresas. Este proyecto se centra en el campo de análisis de datos aplicadas a las empresas Pymes situadas en Guayaquil en donde se combina la ciencia de datos, finanzas de la empresa y gestión de riesgos.

El alcance de este periodo tiene como una investigación que se realizará tomando los datos de un rango de 5 años, considerando en cuenta que tomamos información desde los años 2021-2025, lo cual permitirá analizar la evolución de las empresas durante y después de eventos económicos muy relevantes que pasaron en el país, como la pandemia del COVID 19, paros nacionales y conflictos internos que hubo en el Ecuador, a su vez medidas arancelarias que se impusieron en el mercado global y la recuperación del comercio exterior ecuatoriano.

En relación con la delimitación geográfica ésta se extiende a la ciudad de Guayaquil y todas las empresas cuya actividad económica es la importación de bienes y servicios. Es necesario indicar que este estudio no abarcará empresas de otro sector, ni aquellas que están clasificadas como micro o grandes empresas.

Este proyecto marcara una transformación principalmente en el sector empresarial, incrementando el uso de modelos de machine learning, en empresas importadoras, con la finalidad de que puedan responder efectivamente ante los cambios, con optimización y fortaleza competitiva.

OBJETIVOS

Objetivo General:

Desarrollar un modelo de clasificación para pronosticar quiebras de pymes importadoras en Guayaquil y así contribuir a la prevención de crisis financieras y la sostenibilidad del negocio.

Objetivos Específicos:

- Indagar en la literatura existente sobre aprendizaje automático y su aplicación en la predicción de quiebras empresariales, con el fin de fundamentar teóricamente el desarrollo del modelo predictivo para PYMES importadoras.
- Desarrollar y preparar un modelo clasificatorio mediante técnicas de aprendizaje automático supervisado, utilizando árboles de decisión y random forest, aplicados a datos financieros de PYMES importadoras de Guayaquil, para evaluar la probabilidad de riesgo de quiebra.
- Analizar los resultados obtenidos del modelo predictivo, que favorezcan a la prevención de riesgos financieros en las PYMES del sector importador.

Capítulo I: Fundamentación Teórica

MARCO TEÓRICO

Las PYMES en el contexto económico ecuatoriano

Las PYMES son una fuente clave de dinamización económica, ya que contribuyen al empleo formal, la innovación y el comercio. En Ecuador, estas empresas representan más del 90% del total de unidades productivas y generan cerca del 60% del empleo (UTPL, 2024). De acuerdo con la Superintendencia de Compañías (2025) las PYMES dominan gran parte del comercio, la manufactura y los servicios. Esta estimación coloca a las PYMES es una de las principales fuentes de economía en el territorio nacional, debido al impacto que han causado positivamente en el contexto laboral.

El alcance que tienen las PYMES no solo se relaciona al área económica, sino también al aspecto social, pues se consideran actores examinadores en la reducción de la informalidad y la distribución del ingreso (CEPAL, 2025). Sin embargo, es necesario resaltar que aún existe un panorama incierto con relación a los aspectos de desarrollo debido a la vulnerabilidades financieras, las cuales limitan la continuidad y crecimiento económico, generando restricciones a créditos y escaso manejo administrativo profesional.

A continuación, se presenta una tabla donde se muestra la clasificación de las empresas en el Ecuador, por categoría, ventas anuales y números de empleados, estos datos permiten identificar como se estructuran las PYMES actualmente.

Tabla 1
Clasificación de empresas según el tamaño en Ecuador

Categoría	Ventas anuales (USD)	N.º empleados
Microempresa	0 – 100.000	1 – 9
Pequeña empresa	100.001 – 1.000.000	10 – 49
Mediana empresa	1.000.001 – 5.000.000	50 – 199
Grande	Más de 5.000.000	Más de 200

Fuente:(INEC, 2021, p. 6)

La clasificacion presentada permite conocer la categorizacion que tienen las PYMES en el ecuador, conociendo el monto de sus ingresos economicos y cantidad de personal con el que se cuenta para sus operaciones, dicha informacion, facilita el diseño de politicas tributarias y laborales en el pais, tambien la estructura operativa, necesidades y desafios que enfrentan estas empresas principalmente en terminos de competitividad financiera.

Las PYMES importadoras en Guayaquil

Hablando del ámbito local las PYMES importadoras en guayaquil constituyen un papel importante dentro de la económica de esta ciudad, debido al gran movimiento mercantil que se genera. De acuerdo con los datos proporcionados por la Cámara Marítima del Ecuador (2021) Guayaquil es el principal núcleo comercial del país con más del 93,5% de las operaciones marítimas de importación. El rol que cumplen estas importadoras se evidencia en la capacidad para dinamizar la economía particular, abasteciendo a los negocios minoristas, emprendimientos, mayoristas y consumidores finales. Estas empresas contribuyen al crecimiento empresarial al ofrecer alternativas competitivas y novedosas en el mercado ecuatoriano.

Pese a que las PYMES importadoras en Guayaquil poseen una gran capacidad de solvencia, existen ciertas características que las diferencian de otras empresas, y que han causado un decrecimiento en su desarrollo competitivo, estas características son:

- Alta dependencia del tipo de cambio
- Importación de mercancías sujetas a tributos aduaneros
- Elevados costos logísticos y portuarios
- Necesidad de liquidez constante
- Uso frecuente de financiamiento a corto plazo (CAMAE, 2021).

Estas inconsistencias generan un entorno de alto riesgo principalmente en el área financiera, donde se puede evidenciar la quiebra debido a factores internos (contabilidad, inventarios) y externos (inflación internacional, regulaciones). A esto se suman desafíos administrativos como la disponibilidad de financiamiento, la gestión eficiente de inventarios y la planificación de la demanda para evitar sobrecostos por almacenamiento o faltantes de stock (Andino et al., 2022).

Es así que, las PYMES en guayaquil tienen un papel crucial en la economía regional, ya que originan el flujo de bienes, promueven la innovación comercial y generan empleo. No obstante, para conservar su sostenibilidad y crecimiento, es necesario que fortalezcan sus capacidades de gestión logística, financiera y estratégica ante un entorno global versátil.

Riesgo financiero e insolvencia empresarial

Cuando hablamos de riesgo financiero se lo relaciona directamente con la probabilidad de que una empresa no pueda cumplir con sus obligaciones financieras en un determinado tiempo, esta acción puede verse influenciado por diversos factores, siendo uno de ellos la insuficiencia, capacidad de liquidez o mala gestión del capital (Vaca Sigüeza & Orellana Osorio, 2020). En el ámbito de las PYMES ecuatorianas, este tipo de riesgo suele ser más evidente debido a limitaciones estructurales como poco acceso a financiamiento, altos costos operativos y ausencia de sistemas formales de

planificación financiera, lo cual incrementa la posibilidad de caer en insolvencia.

La (Corte Nacional de Justicia , 2021) manifiesta que, cuando ocurre esta insolvencia la empresa deja de tener la capacidad real de pago. Es decir, una empresa no cuenta con los activos corrientes suficientes para cubrir sus obligaciones, y su patrimonio neto se vuelve negativo a esto se declara como insolvencia económica o quiebra técnica. Esta situación representa la antesala del colapso empresarial y genera un alto nivel de incertidumbre tanto para acreedores como para la continuidad operativa del negocio, especialmente en sectores vulnerables como el comercio exterior.

En el aspecto técnico contable el riesgo financiero se refleja mediante la siguiente representación:

$$Activos < Pasivos$$

Es decir, la empresa no cuenta con flujos suficientes para cubrir sus deudas. En el caso de las PYMES de Guayaquil, cada día este problema se vuelve más intenso, debido a la exposición a factores externos como la variación del tipo de cambio, costos logísticos, lentitud aduanera y fluctuación de la demanda, dichos elementos afectan la liquidez y deterioran la situación económica (Malakauskas y Lakštutienė, 2021).

Es importante resaltar que la insolvencia en una organización no es un evento repentino, sino el resultado de la acumulación progresiva de diversas debilidades financieras. Por lo cual es necesario identificar a tiempo mediante indicadores claves sobre la situación real de la empresa, lo cual resulta esencial para predecir la quiebra y prevenir de algún modo el deterioro del negocio (Ludeña Dávila y Tonon Ordóñez, 2021).

Diversos estudios realizados por (Suescum et al., 2025) en base a las PYMES en Latinoamérica revelan que aquellas que se dedican a actividades de importación son más susceptibles y presentan mayores niveles de riesgo financiero, esta situación ocurre debido a la dependencia de proveedores extranjeros y la variabilidad de los mercados internacionales, lo que incrementa la probabilidad de incumplimiento financiero.

Debido a esta realidad se considera importante desarrollar modelos predictivos basado en datos reales, que permitan identificar señales o alertas a tiempo de insolvencia o riesgos financieros con la finalidad de fortalecer la estabilidad económica de las PYMES importadoras de Guayaquil.

Modelos tradicionales de predicción de quiebras

El análisis del riesgo financiero y la insolvencia empresarial no solo se basa en el estudio de indicadores contables, sino también en la aplicación de modelos especializados que permiten anticipar escenarios de quiebra. En este contexto, los modelos tradicionales de predicción de quiebras y, más recientemente, las técnicas de Machine Learning, se han consolidado como herramientas clave para evaluar la estabilidad económica de las organizaciones, especialmente en sectores vulnerables como las PYMES importadoras de Guayaquil (Jardón et al., 2024).

Entre los más representativos se encuentran:

- **Modelo Z-Score de Altman (1968)**: Evalúa el riesgo de quiebra mediante un conjunto de ratios financieros relacionados con liquidez, rentabilidad y estructura de capital. Altman demostró que las empresas con baja capacidad de pago y elevado endeudamiento tienen mayor probabilidad de insolvencia (Scherge et al., 2018 p.5).
- **Modelo de Beaver (1966)**: Considera el análisis univariado de indicadores como el flujo de efectivo sobre deuda total para detectar señales tempranas de quiebra (Vargas Charpentier, 2020).
- **Modelo Logit-Probit de Ohlson (1980)**: Utiliza técnicas econométricas para estimar la probabilidad de quiebra mediante variables financieras como tamaño de la empresa, liquidez y rentabilidad (Scherge et al., 2018 p.5)

Estos modelos siguen siendo vigentes debido a su capacidad para interpretar de manera sencilla los factores que pueden conducir a una empresa a la insolvencia (Laguillo Diaz, 2020). Sin embargo, presentan limitaciones porque dependen únicamente de datos contables y su precisión

disminuye en entornos altamente cambiantes, como el comercio internacional donde operan las PYMES importadoras.

Justificación de clasificación para el problema de predicción de quiebra

La clasificación de un problema o predicción de quiebra es un tema naturalmente binario, debido a que muchas empresas se autodenominan como solventes o insolventes. Según (Molina et al., 2023) los clasificadores supervisados superan a los modelos tradicionales en precisión, especialmente en escenarios donde las variables tienen relaciones no lineales, como ocurre en las PYMES importadoras ecuatorianas.

Figura 1

Proceso de ciencia de datos aplicado a problemas financieros

Inteligencia Artificial (IA)

Según (Riquelme y Pereira, 2024) la IA no solo optimiza procesos mediante el procesamiento avanzado de datos, sino que también impulsa la creación de ventajas competitivas sostenibles y fortalece la resiliencia organizacional. Esto se debe a que los sistemas inteligentes permiten anticipar riesgos, adaptar estrategias y maximizar la eficiencia, resultados que son particularmente notables en sectores como el de las PYMES importadoras, las cuales se encuentran expuestas a factores externos, y por ende se exige una toma de decisiones ágil, informada y basada en evidencia.

En el contexto empresarial, la IA actúa como una tecnología habilitadora que transforma la forma en que las organizaciones procesan la información, optimizan sus recursos y responden a los cambios del entorno. En el ámbito ecuatoriano, la IA toma un rol estratégico debido a que contribuye

a reducir la incertidumbre y a mejorar la competitividad frente a escenarios económicos etéreos.

La capacidad que posee la IA proporciona diagnósticos más precisos y automatizados principalmente en tareas operativas, que facilita a las empresas responder más rápido a las fluctuaciones financieras, logísticas y comerciales que caracterizan a este tipo de economías.

Figura 2
Inteligencia Artificial

Machine Learning aplicado a la predicción de quiebras

Los modelos de Machine Learning no requieren supuestos estadísticos estrictos y pueden identificar patrones ocultos en los datos. Según (Toapanta et al., 2024) esta acción presenta técnicas que permiten:

- Manejar grandes volúmenes de datos
- Analizar relaciones no lineales
- Combinar variables financieras y cualitativas
- Optimizar predicciones mediante entrenamiento iterativo

El autor (Bonilla Sierra, 2025) manifiesta que el machine learning o también conocido como aprendizaje automático es un subconjunto de las ciencias computacionales, como es también una rama de la inteligencia artificial, que logra facilitar los sistemas para aprender sobre los datos, en ocasiones logra manejar grandes cantidades de datos para crear sistemas automáticos que de forma automática aprenden y tienen las tareas de clasificar, resolver y predecir escenarios críticos o eventos, y comportamientos

sin que sea necesario la práctica de un ser humano, todo esto se logra por la creación y entrenamiento de algoritmos con extensas bases de datos, pudiendo identificar patrones y como resultado el análisis completo de los datos.

Por su parte, (BBVA, 2024) considera que el aprendizaje automático por medio de la utilización de algoritmos lo que permite a las computadoras mejorar su capacidad de aprender, analizar y poder relacionar los patrones de una gran base de datos que pueda llegar a predecir escenarios con estos datos que puedan permitir a las personas un mejor control en su toma de decisiones, por lo que la Inteligencia Artificial, dota a las máquinas la capacidad de ejecutar tareas autónomas sin la necesidad de tener una previa programación ni una supervisión.

Por su parte (Erazo Peña, 2024) indica que, los algoritmos empleados cuando se aplican en el Machine Learning se crean en base a los datos recopilados, en donde se analizan y se calculan dentro de sus capas, y mientras tenga más datos que procesar almacenados en el sistema, estos serán más complejos, pero a la vez tendrán un mayor porcentaje de efectividad. Entonces cuando se llegan a los resultados esperados, serán mucho más exactos e irán mejorando con el tiempo

La predicción de quiebras ha sido históricamente un campo de estudio relevante dentro de las finanzas corporativas, debido a su importancia para la gestión del riesgo empresarial y la prevención de crisis financieras. Según (Monelos et al., 2021) los modelos clásicos como el modelo Z-Score de Altman (1968), el modelo Logit de Ohlson (1980) y el análisis discriminante lineal han sido ampliamente utilizados por su simplicidad y capacidad para evaluar la insolvencia a partir de indicadores financieros. No obstante, estos modelos presentan limitaciones importantes al asumir relaciones lineales entre variables y requerir supuestos estadísticos estrictos, como normalidad, independencia y no multicolinealidad, condiciones que rara vez se cumplen en los datos financieros de las PYMES importadoras ecuatorianas.

Los algoritmos empleados en machine learning evolucionan a medida que procesan más datos. Según (Erazo Peña, 2024) mientras mayor sea el

volumen de información disponible para entrenamiento, más complejos se vuelven los modelos, pero también aumentan su exactitud, capacidad de generalización y eficiencia predictiva. Esto resulta esencial para las PYMES importadoras, que suelen trabajar con información variada, fluctuante y afectada por factores externos como transporte, aranceles y ciclos de demanda internacional.

(Plaza, 2024) explican que el Machine Learning permite integrar no solo variables financieras, sino también factores externos como fluctuaciones del tipo de cambio, costos logísticos, comportamiento del mercado internacional y tendencias de consumo, lo cual lo hace especialmente útil para empresas que dependen de importaciones.

En el contexto de las PYMES importadoras de Guayaquil, el uso de Machine Learning puede ser altamente beneficioso porque:

- Permite identificar el impacto del tipo de cambio en la liquidez.
- Analiza el comportamiento de inventarios importados frente a la demanda local.
- Detecta riesgos derivados del incremento de tarifas, fletes o retrasos en aduanas.
- Mejora la planificación del capital de trabajo ante ciclos de pago más largos. (Altamirano Pérez, 2025)**.**

En el contexto empresarial, el Machine Learning puede procesar variables financieras, operativas y de contexto simultáneamente, para estimar con mayor precisión la probabilidad de quiebra o distress financiero existen diversos métodos como: redes neuronales, Radom Forest, boosting, modelos híbridos (Dasilas y Rigani, 2024). Estos métodos ofrecen múltiples predicciones de forma más confiable y fuerte a diferencia de otros modelos tradicionales.

Importancia, ventajas y resultados de aplicar estos modelos
De acuerdo al estudio presentado por (Leyva y Gómez, 2021) aplicar modelos de predicción de quiebras es esencial para anticipar problemas de liquidez,

tomar decisiones estratégicas y garantizar la permanencia en el mercado. En un entorno vulnerable como el sector importador, estos modelos ayudan a prevenir crisis financieras generadas por factores externos.

Ventajas principales:

- Mayor precisión en la estimación del riesgo financiero.
- Detección temprana de insolvencia, lo que facilita acciones correctivas.
- Mejor gestión del capital de trabajo, especialmente útil para empresas con ciclos largos de recuperación del capital.
- Optimización en la toma de decisiones basada en evidencia cuantitativa.
- Facilidad para acceder a financiamiento, ya que se proyecta un menor riesgo (Vargas Charpentier, 2020)

Resultados esperados:

Según (Moloina Ayala y Piedra Aguilera, 2024) al aplicar este tipo de tecnología se obtendrán los siguientes resultados:

- Reducción de la probabilidad de quiebra.
- Mejora en la estabilidad financiera de las PYMES importadoras.
- Incremento de la rentabilidad y eficiencia operativa.
- Mayor sostenibilidad empresarial en mercados dinámicos.

Ciencia de datos y analítica predictiva

En la actualidad la ciencia de datos es uno de los principales campos de estudio para análisis financiero, debido a la capacidad de transformar grandes cantidades de datos en conocimiento útil para decisiones estratégicas. Según (Pérez et al., 2025), define a la ciencia de datos como la combinación estadística, programación y métodos computacionales para identificar patrones complejos en los datos y predecir comportamientos futuros.

Desde el ámbito empresarial, la analítica predictiva permite anticipar situaciones como el riesgo predictivo, incumplimiento financiero o posibilidad de insolvencia, constituyéndose así en una herramienta clave para las PYMES que operan en entornos cambiantes como el del comercio exterior.

En este mismo sentido (Broby, 2022) señala que la analítica predictiva facilita la estimación probabilística de eventos futuros, a través de modelos estadísticos y matemáticos, con la aplicación de algoritmos de aprendizaje autónomo, lo cual incrementa la precisión del análisis financiero.

Para las PYMES importadoras de Guayaquil, estos mecanismos son verdaderamente relevantes, ya que trabajan con variaciones principalmente en el tipo de cambio, costos de logística, rotación de inventario, y riesgo de créditos. Cuando se integra el análisis predictivo en los procesos financieros se mejora y detecta a tiempo la quiebra empresarial fortaleciendo la gestión del riesgo, permitiendo que las empresas tomen decisiones acertadas en base a sus datos reales y no únicamente en proyecciones subjetivas.

Proceso de ciencia de datos aplicado a problemas financieros

Aplicar la ciencia de datos en las finanzas requiere un proceso estructurado que permite la transformación de información sin filtro en modelos predictivos confiables.

Según (Broby, 2022) este proceso incluye las siguientes fases:

- **Recolección de datos**

  Se basa en la obtención de información relevante proveniente de los estados financieros, registros contables, movimientos bancarios, indicadores macroeconómicos y de la base de datos administrativa. Los datos recolectados deben ser de calidad ya que de eso dependerá la efectividad del modelo de predicción (Broby, 2022).

- **Limpieza**

  Uno de los procesos más significativos es la limpieza, la misma que incluye la eliminación de valores nulos, también consiste en la corrección de errores o detectar outliers mediante la estandarización de formatos. Es importante mencionar que la mayor parte del tiempo

se utiliza en la limpieza, debido a que los datos financieros suelen contener inconsistencias que afectan la precisión del modelo. (Koukaras & Tjortjis, 2025).

- **Selección de variables**

  Este proceso consiste en identificar las variables más relevantes manifestar el riesgo de insolvencia, como liquidez, endeudamiento, rentabilidad e indicadores operativos. De acuerdo con Shetty et al. (2022) una adecuada selección de variables mejora la interpretabilidad e impide el sobreajuste.

- **Modelado**

  Es la etapa en la que se construyen los algoritmos de predicción de quiebra. Incluye técnicas como regresión, árboles de decisión y métodos ensamblados como Random Forest. Trujillo et al. 2025 señalan que los modelos de machine learning han demostrado mayor precisión que los métodos tradicionales en escenarios financieros complejos.

- **Evaluación**

  Este proceso consiste en medir el desempeño del modelo a través de métricas como exactitud, precisión, sensibilidad, y matriz de confusión. Según Uzcátegui et al. (2024) la evaluación con datos de prueba es decisiva para garantizar la generalización del modelo en casos reales.

Árbol de decisión

Los árboles de decisión son modelos de clasificación que segmentan el espacio de datos mediante reglas en forma de “si-entonces”, creando una estructura tipo árbol donde cada nodo representa una prueba sobre una variable y cada hoja una clase de (quiebra / no quiebra) (Scikit Learn, 2025). Este algoritmo es especialmente útil por su interpretabilidad: los gestores pueden ver fácilmente qué condiciones llevan a una predicción de riesgo. Técnicamente usan criterios como Gini o Entropía para seleccionar la mejor división y requieren poda para evitar sobreajuste.

Figura 3
Árbol de decisión

Son modelos que estructuran decisiones en forma de árbol:

$$Impureza\ (Gini) = 1 - \sum_{i=1}^{n} p_i^2$$

Donde p_i es la proporción de un grupo (quiebra/no quiebra).

Proceso de partición y criterios:

- **Índice Gini:** Mide la impureza de los nodos al clasificar los datos; mientras menor es el valor, mayor pureza.
  Es el criterio más usado en clasificación financiera (IBM, 2025).
- **Entropía / Ganancia de información:** Determina qué variable aporta más información al dividir los datos.
  La ganancia de información selecciona atributos clave para identificar riesgo (Yizinet, 2022).

Estructura de un árbol de clasificación

1. **¿Qué es el accuracy?**

Indica la proximidad que tiene una medición o predicción que se realice de un resultado real o verdadero, lo cual se refiere a la propiedad de ser correcto y sin errores y se mide siendo la relación de los aciertos con respecto a todos

los intentos, además, es muy relevante en IA, en análisis de datos y en pruebas debido a que permite validar la fiabilidad que tiene un sistema o un resultado.

Fórmula simple:

$$\boldsymbol{accuracy} = \frac{numero\ de\ predicciones\ correctas}{total\ de\ predicciones}$$

Profundidad, nodos, ramas y hojas

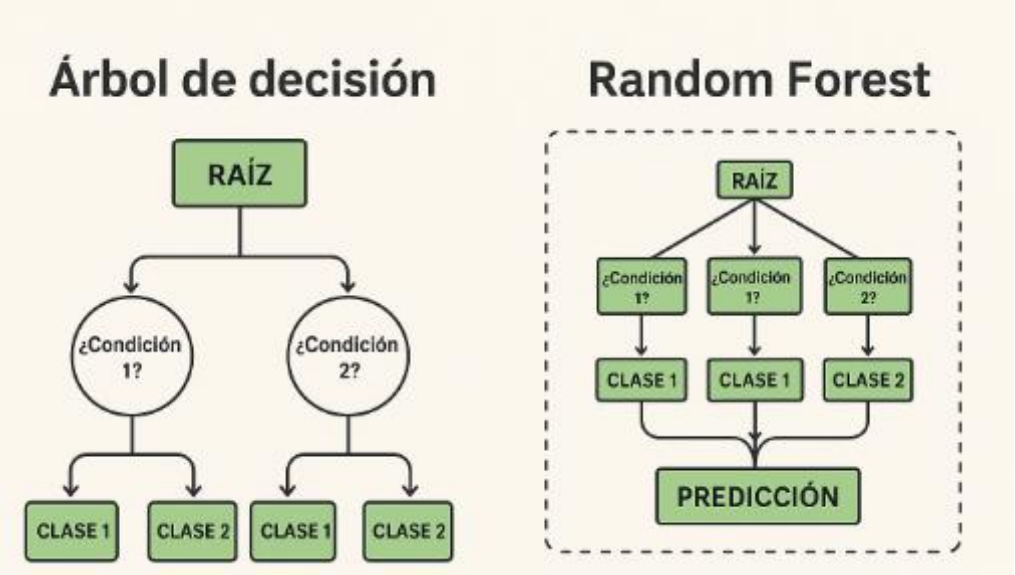
Según (Dunkley Hickin, 2021) por medio de la profundidad se puede indicar la cantidad de divisiones que realiza el árbol; los nodos representan pruebas sobre variables; las ramas, decisiones derivadas; y las hojas, las predicciones finales, tener el control de la profundidad para evitar sobreajustes.

En diferentes estudios sobre PYMES y empresas pequeñas, los árboles han sido de buena utilidad ya que son herramientas de diagnóstico y fuente de reglas adicionales, aunque su precisión podría llegar a ser inferior a la de modelos ensamblados si no se controlan la varianza y la complejidad del árbol.

- **Random Forest**

Random Forest es ideal para combinar variables contables (liquidez, rentabilidad), operativas (rotación de inventarios) y externas (tipo de cambio, fletes) y obtener una predicción robusta para PYMES importadoras en Guayaquil.

Figura 4
Random Forest



Concepto de bagging (bootstrap aggregating)

Según (Marquez, 2022), el bagging consiste en crear múltiples subconjuntos de datos mediante muestreo con reemplazo. Al ejecutar este procedimiento se reduce la varianza y mejora la estabilidad del modelo.

RF utiliza múltiples árboles y combina sus predicciones:

$$RF = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i$$

Donde N es el número de árboles.

Conjunto de árboles independientes

(Marquez, 2022) manifiesta que Random Forest construye múltiples árboles independientes y combina sus predicciones. Este enfoque mejora la precisión al capturar diferentes patrones del mismo conjunto de datos.

Cálculo del voto mayoritario en clasificación

Con relación a clasificación, cada árbol predice una clase, y el modelo final elige la clase más votada. destacan que este mecanismo mejora la estabilidad del modelo (Marquez, 2022).

Cómo reduce el overfitting

Para reducir el overfitting se debe combinar árboles simples, Random Forest el mismo que consiste en la disminución de la varianza y evita que el modelo se ajuste demasiado a los ruidos de los datos. Según Vilema Lara (2022) este mecanismo es uno de los métodos más robustos para información financiera.

Importancia de variables (feature importance)

Para la aplicación correcta de las variables en algoritmos es necesario determinar cuáles son aquellas que más aportan a la predicción. (Medina Merino y Ñique Chacón, 2020) señalan que este indicador permite identificar los factores financieros más determinantes en la quiebra, como liquidez, solvencia o rotación de inventarios.

Qué es Random Forest para regresión

Un Random Forest es un conjunto (o "bosque") de árboles de decisión. En el caso de la regresión, cada árbol de decisión predice un valor numérico y luego las predicciones de todos los árboles se promedian para obtener un resultado más estable y preciso.

Funciona bien sin una preparación de datos extensa.

Ventajas:

- Reduce el sobreajuste
- Promediar muchos árboles evita que un solo árbol dicte la predicción final.
- Robusto al ruido

No implica:

- Escalado
- Normalización
- Transformaciones lineales

Desventajas

- Menos interpretabilidad
- No puede explicarse tan fácilmente como un único árbol.
- Más caro computacionalmente
- Entrenar cientos de árboles requiere memoria principal y tiempo

Paso a paso conceptual

Paso 1: Preparar los datos.

- Se divide el dataset en X (variables independientes) e Y (variable a predecir).
- Se suele separar en conjunto de entrenamiento y prueba para evaluar el modelo.

Paso 2: Construir varios árboles de decisión.

1. **Para cada árbol:**

Se toma una muestra aleatoria con reemplazo del dataset de entrenamiento (**bootstrap sampling**). Esto significa que algunos datos pueden repetirse y otros no aparecer en ese árbol. Esto crea diversidad entre los árboles.

2. **Al construir cada nodo del árbol:**
   - En lugar de considerar todas las variables, se selecciona un subconjunto aleatorio de variables.
   - Esto hace que los árboles sean diferentes entre sí y reduce la correlación entre ellos.
3. **Cada árbol se construye hasta:**
   - Una profundidad máxima definida, o
   - Un número mínimo de muestras por hoja, o
   - Otra condición de parada.

Paso 3: Predecir con cada árbol

Cada árbol devuelve un valor numérico para la observación que quieres predecir.

Paso 4: Promediar las predicciones

La predicción final del Random Forest es la media de todas las predicciones individuales de los árboles:

Paso 5: Evaluar el modelo

Para regresión, se utilizan métricas como:

MSE (Mean Squared Error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

N = número de observaciones

Yi= valor real

Yi = valor predicho

RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MAE (Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

R² (R cuadrado)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Evaluación con MSE

¿Qué hace esta línea?

Calcula el **Mean Squared Error**, que se define como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{real} - y_{predicho})^2$$

Es decir:

- Toma la diferencia entre valor real y predicción.
- Lo eleva al cuadrado.
- Promedia todos esos valores.

El MSE mide cuánta distancia hay entre las predicciones y los valores reales.

Marco Conceptual

Sistemas Inteligentes en el Análisis Financiero

Los sistemas inteligentes se relacionan directamente con el uso de la tecnología y la aplicación de inteligencia artificial, cuya finalidad es automatizar diversas actividades y que minimicen la actividad humana, mediante la resolución de problemas complejos.

Según (Herrera Sánchez y Casanova Villalba, 2024) en el sector financiero la incorporación de sistemas inteligentes ha transformado la forma en que las empresas analizan riesgo, proyectan escenarios y previenen quiebras. Los sistemas inteligentes son herramientas computacionales diseñadas para simular capacidades humanas como la toma de decisiones, la detección de patrones y el razonamiento automático.

En el entorno empresarial, estos sistemas favorecen al análisis predictivo, anticipando riesgos de insolvencia por medio de modelos basados en grandes cantidades de datos. En relación al Ecuador, esta aplicación se ha convertido en un componente esencial para el correcto funcionamiento del sistema financiero y para el direccionamiento en la toma de decisiones estratégicas internas de cada organización. En los últimos años este sector se ha visto obstaculizado por grandes retos los cuales son bastantes significativos, principalmente a la hora de incorporar de forma progresiva la tecnología digital y con la necesidad de responder a un entorno empresarial que se transforma con rapidez y exige mayor competitividad (Ocampo Alvarado, 2024).

De esta forma es necesario que los profesionales del área comercial y contable requieran una actualización permanente, por medio de procesos de capacitación interna y replantear sus métodos de trabajo para hacerlos más eficientes y alineados a las nuevas exigencias del mercado.

Enfoques Computacionales para la Predicción de Riesgo Empresarial

En toda organización son necesarios los enfoques computacionales debido a su uso constante en la predicción de riesgos empresariales, financieros y operativos. Es así como los enfoques computacionales para la predicción del riesgo empresarial consisten en la aplicación de técnicas de análisis de datos, modelado estadístico y algoritmos de aprendizaje automático para anticipar la probabilidad de insolvencia, fracaso o quiebra de una organización, a partir de sus datos financieros, operativos o de contexto (Jones, 2023).

Dicho enfoque se vuelve indispensable principalmente para las PYMES importadoras, donde casa riesgo está asociado a la variabilidad, tiempo de importación, costos logísticos y disponibilidad de liquidez. La aplicación de este mecanismo beneficiara a las PYMES importadoras y permitirá detectar a tiempo situaciones complejas que puedan ocasionar paralizaciones en sus operaciones y decrecimiento en su progreso económico.

Modelos predictivos

En PYMES, la aplicación de modelos predictivos ayuda a prever fluctuaciones del mercado, riesgos financieros y variaciones en el tipo de cambio, especialmente en empresas importadoras que están expuestas a factores externos. Según Trujillo et al. (2025) los modelos predictivos son algoritmos que permiten estimar eventos futuros mediante el análisis estadístico y el aprendizaje automático.

Estos modelos, no solo presenten pronósticos de situaciones futuras, sino que también proporcionan datos reales y claves para establecer alertas tempranas y crear estrategias de mitigación frente a panoramas adversos. Como señalan Shmueli y Koppius (2025), los valores provenientes de la predicción se basan en la capacidad de dirigir acciones concretas y que estas mejoren la toma de decisiones, facilitando a que las empresas actúen con mayor prevención y minimicen las dudas congénitas de mercados emergentes.

De esta manera, los modelos preventivos contribuyen a la planificación organizacional, y la optimización de los recursos permiten actuar de manera eficiente ante los riesgos económicos y operativos. Con la finalidad de conservar a las PYMES importadoras en su continuidad comercial en el mercado.

Modelos tradicionales de predicción de quiebras

Antes del surgimiento del aprendizaje automático, las predicciones de quiebra dependían de diversos modelos estadísticos y econométricos, también se realizaba el análisis de la discriminante, modelos *logit/probit,* y

fórmulas basadas en ratios financieros históricos (liquidez, endeudamiento, rentabilidad, entre otros (Molina Panchi y Molina Panchi, 2024).

Estos modelos usan información contable previa para estimar la probabilidad de insolvencia, asumiendo relaciones lineales entre variables y dependencia exclusiva de datos financieros. Con la finalidad de comprobar la situación de las empresas previo a su predicción.

Machine Learning (Aprendizaje Automático)

El Machine Learning (ML) es un conjunto de técnicas computacionales que facilitan la construcción de modelos predictivos a raíz de datos históricos de empresas, sin necesidad de definir reglas explícitas en donde los algoritmos “aprenden" de los datos (Luna et al., 2024).

Inteligencia Artificial

La Inteligencia Artificial conocida también como (IA) comprende un conjunto de técnicas y sistemas diseñados para ejecutar tareas que tradicionalmente requieren de la capacidad cognitiva humana, tales como el análisis de grandes volúmenes de datos, la identificación de patrones complejos, la predicción de eventos futuros y la toma de decisiones basada en información automatizada (Gobierno de España, 2023).

Minería de datos (Data Mining)

(Uriarte del Águila, 2024) definen la minería de datos como un proceso analítico que aplica técnicas estadísticas, computacionales y algoritmos inteligentes para explorar grandes volúmenes de información y detectar patrones significativos, estructuras ocultas y relaciones que no pueden identificarse mediante métodos tradicionales.

(Ruiz Díaz de Salvioni y Armoa, 2023) afirman que, por medio de la minería de datos se pueden alcanzar tendencias que son necesarias para el análisis convencional de los modelos predictivos. Ya que, la minería de datos aplica técnicas para indagar sobre conjunto de indicadores e información útil para obtener conocimiento oportuno.

En las PYMES, esta herramienta mejora la organización, controla y detecta fallas en el sistema económico, desarrollando un rol estratégico, debido a su gran capacidad para analizar toda información relacionada a tipos de cambios económicos, costos de importación, comportamiento de proveedores internaciones, tiempo en la entrega de productos y la demanda de las PYMES importadores de Guayaquil. Estos factores muestran alta volatilidad, y su análisis tradicional puede resultar insuficiente ante la prevención de riesgos. De la misma forma, la minería de datos permite detectar anomalías, tanto en el flujo de caja, como en el incremento de los costos logísticos. Estas detecciones son fundamentales para reducir la exposición a riesgos y evitar futuros escenarios de insolvencia empresarial.

Por lo tanto, la minera de datos no solo aporta significativamente al análisis descriptivo antiguo de una empresa, sino que se convierte en un recurso crítico para la construcción de modelos predictivos, mediante la aplicación de algoritmos como arboles de decisión, random forest o redes neuronales, estos mecanismos por su capacidad pueden mejorar la calidad en las predicciones anticipada de posibles riesgos.

Analítica financiera

La analítica financiera comprende el uso de métodos cuantitativos, estadísticos y computacionales que permiten evaluar de manera rigurosa la situación económica de una empresa y anticipar posibles escenarios financieros desfavorables. De acuerdo con (Cargua Pilco, 2020) esta área se encarga de transformar los datos financieros en información relevante mediante el uso de herramientas matemáticas y modelos analíticos, lo que facilita examinar la estructura de los costos, identificar tendencias, evaluar el rendimiento financiero y detectar señales de deterioro económico antes de que impacten en la operatividad del negocio.

Según UNIR (2023) sostiene que la analítica financiera contribuye a que la toma de decisiones empresariales sea más racional y fundamentada, minimizando los errores derivados de la intuición o percepción subjetiva del gerente. Al basarse en datos objetivos y técnicas avanzadas de análisis, la

analítica financiera se constituye en un apoyo esencial para mejorar la transparencia y precisión de las decisiones empresariales.

Es así que, el aporte de la analítica financiera hacia las PYMES importadoras radica en el suministro de datos precisos y estructurados los cuales incrementan la efectividad del modelo predictivo, y fortalecen la capacidad de las PYMES para anticipar crisis y tomar decisiones proactivas que eviten su quiebra definitiva.

Sistemas de apoyo a la toma de decisiones (DSS)

Almadani et al. (2025) define a los Sistemas de Apoyo a la Toma de Decisiones (DSS, por sus siglas en inglés) como plataformas tecnológicas que integran datos, modelos computacionales, algoritmos analíticos y herramientas de visualización para asistir a los directivos en la selección de alternativas estratégica.

AsImismo, Olavsrud(Olavsrud, 2024) señalan que los DSS son esenciales para transformar datos en acciones concretas y permiten procesar grandes volúmenes de información y convertirlos en recomendaciones accionables, facilitando una comprensión más clara de los problemas empresariales y respaldando decisiones fundamentadas en evidencia.

Las PYMES importadoras constantemente se enfrenta a mercados inestables o altamente sensibles, donde los DDSS se vuelven herramientas críticas esenciales para proyectar escenarios que permitan evaluar riesgos y analizar múltiples opciones antes de tomar una decisión. Ante la predicción de quiebras empresariales, los DSS aportan valor porque funcionan como un puente entre los datos financieros y los modelos avanzados de aprendizaje automático.

La capacidad que poseen los DSS para integrar algoritmos como arboles de decisión o random forest, es positiva debido a que ofrecen alertas tempranas sobre probabilidades de insolvencia, no solo ayudan a interpretar los resultados de predicción, sino que también facilitan su aplicación práctica en las PYMES importadoras de Guayaquil, fortaleciendo su resiliencia y capacidad de respuesta ante escenarios adversos.

Mercados emergentes

Los mercados emergentes se distinguen por presentar un crecimiento acelerado acompañado de una elevada volatilidad económica, condiciones que obligan a las empresas a adoptar herramientas tecnológicas capaces de reducir riesgos y mejorar su capacidad de respuesta (Pérez, 2025). Este tipo de mercado se caracteriza por su crecimiento acelerado y alta volatilidad económica, lo que exige herramientas que mitiguen riesgos y mejoren la adaptabilidad empresarial.

De acuerdo con Chacón et al. (2024) estas economías poseen estructuras cambiantes, incertidumbre institucional y vulnerabilidad a fluctuaciones externas, los mercados internacionales, hace imprescindible el uso de sistemas computacionales que permitan evaluar los posibles impactos antes de que estos ocurran.

PYMES importadoras

Las PYMES importadoras son pequeñas y medianas empresas dedicadas a la compra de productos o insumos desde el extranjero. Su vulnerabilidad ante cambios en el tipo de cambio, costos logísticos, aranceles y tiempos de entrega demanda herramientas avanzadas para gestionar el riesgo (Andino et al., 2022).

Para (Zambrano et al., 2019) este tipo de empresas se requiere sistemas más ágiles de análisis financiero y de planificación para reducir la exposición a factores externos. Esto permitirá que las PYMES obtengan mayor certeza en los análisis predictivos que se realicen, y que se puedan tomar decisiones acertadas en base a su situación.

Transformación digital

Según (Villegas Rojas y Altamirano Freire, 2025) la transformación digital integra herramientas digitales en todos los procesos operativos y de decisiones estratégicas. En mercados emergentes, este proceso permite aumentar la competitividad y mejorar la eficiencia. Cuando se refiere a la transformación digital no solo se hace énfasis en la introducción de dicho

sistema, sino también en las modificaciones estructurales organizacionales y la forma en cómo se generan los valores y oportunidades para las empresas.

En este sentido, la transformación digital se convierte en un factor determinante para reducir brechas económicas y operativas en economías en desarrollo, ya que facilita la automatización de procesos, la optimización del uso de recursos y la generación de nuevas capacidades productivas. Como destacan (McCartney, 2023) las organizaciones que integran tecnologías digitales avanzadas logran una mayor capacidad de respuesta ante entornos cambiantes.

Gestión del riesgo

La gestión del riesgo es el proceso mediante el cual una empresa identifica, evalúa y responde a posibles amenazas que pueden afectar su operación y sus resultados (Martins, 2025).

Para las PYMES importadoras, la gestión del riesgo debe incluir variables como variación del tipo de cambio, retrasos aduaneros y costos inesperados de logística. Según ISO 31000 (2023), la gestión del riesgo debe basarse en análisis sistemáticos y datos cuantitativos, lo cual refuerza la importancia de los enfoques computacionales.

Aprendizaje Automático como Método Predictivo

De acuerdo con (UNIR, 2025) el aprendizaje automático (Machine Learning – ML) es un conjunto de técnicas que permiten que los modelos computacionales aprendan patrones a partir de datos históricos y generen predicciones sin intervención humana directa.

Existen múltiples características de los Machine Learning que son fundamentales para la predicción de riesgos empresariales, como:

- Capacidad de procesar datos masivos
- Manejo de relaciones complejas
- Flexibilidad ante cambios en el entorno económico
- Mejora continua mediante entrenamiento iterativo

Para PYMES importadoras, estas características permiten evaluar cómo factores financieros y logísticos influyen en la probabilidad de quiebra. Detentando a tiempo posibles eventos negativos que pueden perjudicar considerablemente a las organizaciones.

Algoritmos Supervisados para la Predicción de Insolvencia

Para determinar la situación real de una empresa específicamente en términos económicos y de rentabilidad, se requiere de variados estudios y análisis interpretativos. Sin embargo, en los últimos tiempos es inevitable mencionar el surgimiento de algoritmos debidamente comprobados que facilitan la predicción de insolvencia en las organizaciones.

(Franco Gómez, 2023) describe que en la investigación financiera se utilizan algoritmos supervisados debido a que existe una variable objetivo-conocida (empresa quebrada o empresa solvente) por lo que se destacan que estos modelos son adecuados para clasificación binaria, como en el caso de la predicción de quiebra.

Entre los algoritmos más utilizados están:

- **Árboles de decisión (Decisión Tree Classifier)**

  Rivadeneira et al. (2022) define un árbol de decisión como un modelo que divide datos en varios segmentos por medio de reglas basadas en valores que contienen las variables de un determinado caso. Este método es interpretable ya que refleja las decisiones tomadas y permite la aplicación de condiciones financieras específicas.

- **Random Forest**

  Es un método ensemble que construye muchos árboles de decisión sobre subconjuntos bootstrap de los datos y promedia (o vota) sus predicciones (Ahmed at al., 2024). Random Forest también proporciona medidas internas de importancia de variables lo que ayuda a identificar cuáles indicadores financieros o logísticos explican mejor la quiebra.

Limitaciones de los Modelos Tradicionales de Predicción de Quiebras

Isaac y Caicedo (2023) afirman que los modelos de Altman (1968), Beaver (1966) y Ohlson (1980) fueron pioneros en el análisis de insolvencia empresarial. Sin embargo, investigaciones recientes demuestran que presentan limitaciones para los contextos actuales. Según (Romero Espinosa, 2023) estos los modelos tradicionales presentan las siguientes limitaciones:

- Tienen capacidad limitada para manejar datos no lineales
- No se adaptan bien a sectores con alta volatilidad
- No incorporan información cualitativa o variables externas
- Reducen la capacidad predictiva en PYMES

Por ello, surge la necesidad de emplear modelos más robustos como los basados en ML.

Marco Legal

En el contexto legal de esta investigación se tomará como referencia la Superintendencia de Protección de Datos Personales (SPDP) que es la entidad pública encargada de garantizar el cumplimiento de la Ley Orgánica de Protección de Datos Personales en el Ecuador, conforme a lo establecido en el artículo 213 de la Constitución de la República. Su rol resulta especialmente relevante en investigaciones que emplean modelos de datos y técnicas de aprendizaje automático, como la presente tesis orientada a la predicción de quiebras de PYMES importadoras en Guayaquil, debido al uso intensivo de información financiera y administrativa.

El desarrollo de un modelo predictivo basado en árboles de decisión y Random Forest implica la recopilación, procesamiento y análisis de datos financieros históricos, tales como estados de resultados, balances generales, niveles de endeudamiento, liquidez y comportamiento de pagos. De acuerdo con la SPDP (2025), este tipo de información puede considerarse dato personal o dato personal económico ya que permite identificar directa o

indirectamente a una persona natural asociada a la empresa, como socios, representantes legales o responsables financieros, por lo que su tratamiento debe regirse bajo principios de licitud, finalidad, minimización y seguridad.

En el ámbito jurídico la Ley Orgánica de Protección de Datos Personales (LOPDP) es la norma central que regula el tratamiento de datos, misma que menciona en el Artículo 2 “Ámbito de aplicación: La ley se aplica a todo tratamiento de datos personales que consten en cualquier soporte o medio” (2021, p. 5).

Es necesario mencionar que la SPDP, exige que todos los análisis ejecutados mediante procesos automatizados o uso de algoritmos para predecir quiebras se desarrolle con responsabilidad eficiente, y que esta se relacione con toda la estructura planteada en esta investigación. La elaboración de un modelo predictivo que detecte quiebra empresarial aplica fases como la recolección, limpieza, selección de variables, modelado y evaluación, mismos que han sido explicados en el marco teórico, con el único fin de documentar de forma clara cada una de las evidencias que son necesarias y únicamente con fines académicos.

El Plan Nacional de Protección de Datos Personales (PNPDP), emitido por la SPDP, promueve el eje de innovación segura, el cual busca fomentar el uso de tecnologías emergentes, como la inteligencia artificial y el machine learning, siempre que estas sean aplicadas con criterios de transparencia y control, manifestados en el artículo 47 de la (LOPDP, 2021) y en el artículo 66 numeral 19 de la Constitución de la República del Ecuador (2021), los mismos que garantizan que incluso en procesos predictivos nadie pierda el control de su información.

En relación a los árboles de decisión y Random Forest, se considera que estos algoritmos facilitan la interpretación de variables financieras, lo que permite determinar cómo y porque una empresa se considera solvente, insolvente o en posible quiebra, dichos resultados deben estar alineados a las exigencias de la autoridad de control actual.

La SPDP instituye que los responsables del tratamiento de información deben aplicar medidas de seguridad técnicas y organizativas, como la anonimización de los datos y el control de accesos. El estudio desarrollado, muestra que los datos analizados en este contexto deben estar depurados, sin identificadores directos de personas naturales, logrando que los resultados de la observación se ajusten únicamente en la predicción del riesgo empresarial y no en la evaluación individual de personas.

La Superintendencia de Protección de Datos Personales se sujeta claramente con la indagación ejecutada, estableciendo un marco regulatorio basado en la predicción de quiebra y ciencia de datos, garantizando un equilibrio entre la innovación tecnológica y el sostenimiento económico, protegiendo así el modelo presupuestario de las PYMES de Guayaquil.

METODOLOGÍA

En este capítulo se explican los pasos utilizados para crear un modelo de clasificación que prediga quiebras en PYMES importadoras de Guayaquil, empleando árboles de decisión y Random Forest con datos históricos de 2021 a 2025.

Se detalla el proceso completo: adquisición y limpieza de datos financieros, creación de variables, entrenamiento de modelos, validación y análisis de resultados, con criterios de trazabilidad y custodia de la información en cumplimiento de la normativa ecuatoriana.

Tipo de investigación

El análisis es de tipo cuantitativo, porque se ocupa de datos financieros numéricos y examina un modelo de clasificación utilizando parámetros objetivos. De acuerdo con Mendoza (2006) "la investigación cuantitativa es aquella en la que se recopilan y examinan datos numéricos de variables".

Diseño de investigación

La investigación no es experimental, ya que no se alteran las variables; en su lugar, se examinan los registros financieros preexistentes (2021-2025) para formar y evaluar árboles de decisión y Random Forest. De acuerdo con Hernández et al. (2006) "es el que se lleva a cabo sin alterar intencionadamente las variables. Se fundamentos sobre todo en la observación de los fenómenos" (p. 6).

Alcance de la investigación

El alcance se caracteriza por ser correlacional y predictivo, pues se analiza la conexión entre los indicadores financieros y la situación de quiebra, además de que se desarrolla un clasificador para calcularla probabilidad de riesgo. Las categorías fundamentales se presentan en la tipología de alcances: la tipología incluye cuatro tipos de investigaciones: descriptivas, correlacionales, explicativas y exploratorias (Hernández et al., 2006).

Población objetivo

La población de estudio se restringe a los datos históricos internos de importaciones de FEGOCAR en Guayaquil del período 2021-2025, extraídos de su base transaccional y/o estados relacionados de compra (precio, flete, seguro, cantidades, pesos, proveedor, características del producto). Por ser una empresa establecida y con actividad económica en Ecuador, su identificación tributaria corresponde al RUC, instrumento utilizado para censar e identificar a los contribuyentes y las actividades económicas que realizan. Estos registros se utilizan bajo medidas de seguridad de la información y tratamiento de datos, en cumplimiento con la Ley Orgánica de Protección de Datos Personales (Chávez, 2023).

Unidad de análisis

La unidad de análisis es la transacción de importación (registro por operación), susceptible de ser agregada por mes-año o año para fines de elaboración de informes gerenciales, en la estructura del dataset 2021-2025. Operaciones que informan todos los datos para el cálculo de variables derivadas del modelo (precio_unitario, peso_unitario, costo_logistico_ratio), con Precio, Cantidad, Peso neto, Transporte y Seguro válidos, y campos mínimos de trazabilidad como Proveedor, País de Origen, Tipo de mercancía y Modelo. Esta restricción es capaz de medir señales operativas y de compra, sensibles para control de gasto y riesgo por proveedor en el flujo importador que se concentra en Guayaquil como principal punto de entrada al país.

Criterios de inclusión

Se Incluyen compañías que muestrean estados financieros completos por lo menos durante dos períodos en el intervalo de 2021 a 2025, con las cuentas mínimas necesarias para calcular indicadores (activos, pasivos, patrimonio, ventas, costos y resultado del ejercicio). Para evitar situaciones con datos incompletos o incongruencias en la contabilidad que obstaculicen el cálculo de variables predictivas y la asignación de la etiquetade quiebra o no quiebra, se incluyen solamente compañías cuya actividad importadora esté declarada y cuenten con registros consistentes.

Empresa Importadora FEGOCAR

Se trata como caso empresarial a FEGOCAR, empresa del sector importador en Guayaquil, con transacciones históricas de compra internacional y las variables características de una importación (precio de compra, flete/transporte, seguro, cantidades, peso, país de origen, proveedor). En el dataset se codifican categorías de producto y atributos relacionados con mercado automotriz (ej., repuestos, accesorios, modelo de vehículo), lo que posibilita la creación de indicadores unitarios y ratios de costo logístico para control operativo de margen y gasto total.

Como proceso, este tipo de firma se relaciona con proveedores externos, costos logísticos, variación por mezcla de mercancía, por lo que el riesgo se puede asociar con prácticas de gestión de riesgo organizacional enfocadas en la identificación, análisis y monitoreo de la exposición. Con esta acotación, la metodología y los resultados del modelo se atan a definiciones internas de FEGOCAR: priorización de revisión, alertas por transacción, seguimiento por proveedor y trazabilidad de compras en el periodo 2021–2025.

Análisis exploratorio de datos

En esta fase metodológica se realizó un análisis exploratorio de datos con el fin de verificar la calidad del conjunto de registros, describir el comportamiento de las variables y definir criterios previos al entrenamiento de los modelos.

El procedimiento incluyó revisión de estructura y tipos de datos, conteo de valores faltantes, cálculo de medidas de tendencia central y dispersión para variables numéricas, detección de valores extremos mediante percentiles y gráficos tipo histograma y boxplot, además de estimación de asociaciones entre variables cuantitativas con correlación de Spearman.

Estas actividades permitieron depurar inconsistencias, construir métricas derivadas (precio_unitario, peso_unitario, costo_logistico_ratio) y establecer una base cuantitativa para la selección de predictores y la interpretación posterior de la clasificación de riesgo.

Tratamiento y control de calidad de datos

```
#Valores Nulos
df.isnull().sum()
```

Los valores faltantes fueron cuantificados, uno por variable. Un tratamiento se definió para evitar cualquier sesgo o error de cálculo.

Examinar valores ausentes es fundamental para garantizar la integridad del conjunto de datos, antes del análisis y el modelado subsecuentes. Datos faltantes, podrían distorsionar promedios, crear gráficos incompletos y, lo que es peor, causar errores en el entrenamiento de modelos.

Su identificación precisa el tratamiento a seguir, incluyendo la eliminación de filas o columnas, la imputación con la media, la mediana o la moda, permitiendo también justificar exclusiones en el informe final. Un manejo meticuloso de estos valores mejora la exactitud de los resultados, reduce los posibles sesgos y facilita, además, la documentación clara del proceso de preparación de los datos.

Selección de variables numéricas para estadística descriptiva

df_numericas=df.select_dtypes(include=['int64','float64']) para filtrar variables cuantitativas.

Se separaron variables numéricas para aplicar medidas de tendencia central y dispersión; filtra el DataFrame y se queda solo con columnas cuantitativas (números), estos son importante porque:

- Las medidas estadísticas solo tienen sentido en datos numéricos
- Evita errores con variables categóricas (texto)

Cálculo de medidas descriptivas (tendencia central y dispersión)

```
# Calcular medidas de tendencia central y dispersión
medidas = pd.DataFrame({
    'Media': df_numericas.mean(),
    'Mediana': df_numericas.median(),
    'Moda': df_numericas.mode().iloc[0],
```

- **Media**: promedio general
- **Mediana**: valor central (menos afectada por outliers)
- **Moda**: valor más frecuente

Se calcularon medidas para caracterizar distribución y variabilidad, además de cuartiles e IQR para lectura resistente a extremos.

La aplicación concertada de la media y la mediana facilita la evaluación de la forma que adopta la distribución de una variable. Si los valores son próximos, usualmente la distribución denota mayor simetría.

Por otro lado, al distanciarse la media de la mediana, frecuentemente emerge la asimetría producto de valores aberrantes o la aglomeración de datos hacia un extremo del rango. Esta comparación contribuye a la interpretación más precisa de los descriptivos y al seleccionar medidas con mayor representatividad ante la presencia de valores atípicos.

```
'Desviación estándar': df_numericas.std(),
'Varianza': df_numericas.var(),
'Mínimo': df_numericas.min(),
'Máximo': df_numericas.max(),
'Rango': df_numericas.max() - df_numericas.min(),
```

- **Desviación estándar**: cuánto se alejan del promedio
- **Varianza**: dispersión al cuadrado
- **Rango**: distancia total entre el mínimo y el máximo
- **IQR**: dispersión central, robusta a outliers

La desviación estándar la varianza el mínimo el máximo y el rango nos dejan cuantificar cómo se dispersan los datos además de contrastar la estabilidad entre las variables. Una dispersión alta muestra una volatilidad superior con posibles operaciones poco comunes mientras que una dispersión baja revela un comportamiento más constante. El comparar mínimos y máximos a veces es útil para encontrar fallos o registros fuera de lugar antes de entrenar los modelos.

```
'Q1 (25%)': df_numericas.quantile(0.25),
'Q2 (50%)': df_numericas.quantile(0.50),
'Q3 (75%)': df_numericas.quantile(0.75),
'IQR': df_numericas.quantile(0.75) - df_numericas.quantile(0.25)})
```

Esto permite:

- Ver valores extremos
- Construir boxplots
- Detectar posibles outliers

Para las variables numéricas del conjunto de datos, se calculan medidas de tendencia central y dispersión también, eh. Esto abarcó la media, la mediana, y la moda, incluso la desviación estándar, la varianza y el rango intercuartílico, por ejemplo.

Inclusión y exclusión de columnas y desarrollo de histogramas

```
# Columnas a excluir
excluir = ['Año de fabricacion', 'Cantidad', 'precio_unitario', 'peso_unitario']
```

Se establecen variables no deseadas para graficar en los histogramas. Metodológicamente, esto ocurre: columnas especificas ofrecen poca información o "aplastan" la escala. Pongamos un ejemplo: Año de fabricación es discreto, Cantidad puede estar dispersa, y precio_unitario/peso_unitario tienen valores extremos, dominando la visualización. Excluyéndolas, se favorece la interpretación de las variables financieras en la distribución.

```
# Seleccionar solo columnas numéricas permitidas
df_hist = df.select_dtypes(include='number').drop(columns=excluir
```

Esta línea filtra el DataFrame, conservando únicamente columnas numéricas *int* y *float* después borra aquellas señaladas en excluir. *df_hist* el producto final, un conjunto variables a usar en el histograma. Es decir, en terminos simples, esto es como una receta, solo se grafican variables representables en histogramas que facilitan el diagnóstico financiero.

```
# Histograma
df_hist.hist(figsize=(14,10), bins=20)
plt.suptitle("Distribución de variables financieras", fontsize=16)
plt.show()
```

df_hist. hist(.) produce un histograma para cada variable seleccionada, usando veinte intervalos bins. Se observa la concentración de valores, asimetría, y si existen colas muy largas, *figsize* determina el tamaño para garantizar que todos los gráficos se puedan leer fácilmente. *plt. suptitle(.)* añade un título general para la figura, *plt. show()* muestra el resultado final. En análisis financiero, estos histogramas sirven para encontrar variables con distribuciones sesgadas, valores muy fuera de lugar y rangos preponderantes; esto da información útil, para justificar percentiles, reglas de riesgo, y otras transformaciones importantes antes de entrenar el modelo.

Correlación entre variables

```
#Correlación entre variables
plt.figure(figsize=(12,8))
sns.heatmap(df.select_dtypes(include=np.number).corr(), annot=True,
cmap="coolwarm", fmt=".2f")
plt.title("Matriz de correlación")
plt.show()
```

plt.figure(figsize=(12,8))

*C*rea una figura donde se dibujará el gráfico, mientras que *figsize* define el tamaño (ancho, alto) para que el heatmap sea legible.

sns.heatmap(

Este comando llama a *seaborn* para crear un mapa de calor (heatmap); este tipo de gráfico es ideal para visualizar correlaciones.

df.select_dtypes(include=np.number)

Selecciona solo las variables numéricas del DataFrame; evita errores y asegura que la correlación sea válida

.corr()

Calcula la matriz de correlación de Pearson.

Los valores van de:

- -1 → correlación negativa fuerte
- 0 → sin relación
- 1 → correlación positiva fuerte

annot=True

Muestra el valor numérico exacto de la correlación dentro de cada celda.

cmap="coolwarm"

Define la paleta de colores:

- Rojo =correlación positiva fuerte
- Azul= correlación negativa fuerte
- Colores suaves = correlación débil

fmt=".2f"

Da formato a los números para que se muestren con 2 decimales.

plt.title("Matriz de correlación")

Agrega un título descriptivo al gráfico.

plt.show()

Muestra el gráfico en pantalla.

Determinar relación entre variables vs riesgo operativo

```
# Relación variables vs riesgo operativo
for col in df.columns:
    # Excluir variable objetivo y la predicción
    if col not in ['riesgo_alto', 'pred_dt']:

        if df[col].dtype in ['int64', 'float64']:
            sns.boxplot(x='riesgo_alto', y=col, data=df)
            plt.title(f"{col} vs Riesgo Operativo")
            plt.show()
        else:
            # Para variables categóricas
            if len(df[col].unique()) < 50:
                plt.figure(figsize=(10, 6))
                sns.countplot(
                    x=col,
                    hue='riesgo_alto',
                    data=df,
                    palette=[COLOR_LOW, COLOR_HIGH]
                )
                plt.title(f"Distribución de {col} por Nivel de Riesgo")
                plt.xticks(rotation=45, ha='right')
                plt.tight_layout()
                plt.show()
```

Este bloque es útil; se emplea para contrastar cada atributo del conjunto de datos con la etiqueta *riesgo_alto* y generar gráficos que revelen cómo las variables distinguen las transacciones de riesgo bajo y alto. Esencialmente, esto forma parte del análisis exploratorio antes del entrenamiento, facilitando la comprensión de si las clases se distinguen en magnitudes numéricas o categorias.

El código itera sobre todas las columnas (*for col in df.columns),* excluyendo *riesgo_alto* y *pred_dt* para evitar la gráfica de la variable objetivo o una salida del modelo. Si la columna es numérica (*int64 o float64*), entonces, genera un *boxplot* con *riesgo_alto* en el eje X y la variable en el eje Y; esto expone la mediana, dispersión, y valores extremos por clase. En caso de columna categórica, inicialmente comprueba que no tenga demasiadas categorias (< 50) para evitar gráficos ilegibles, y posterior, genera un *countplot* con barras por categoría y colores separados por *riesgo_alto*, permitiendo identificar si ciertas categorías son más frecuentes en riesgo alto.

Modelo supervisado de clasificación con Árbol de Decisión y Random Forest

El modelo base de la programación es la clasificación supervisada, en la que los indicadores financieros son las entradas y la variable objetivo es binaria (quiebra/no quiebra). El Árbol de Decisión crea reglas del tipo "si... entonces" y organiza las particiones en una estructura arbórea; cada hoja asigna una clase y cada nodo evalúa una variable. Al seleccionar cada división, se utilizan medidas de impureza como la entropía o Gini, lo que posibilita la creación de reglas interpretables desde los datos (Balcan y Sharma, 2025).

Random Forest, un método ensamblado que combina las predicciones de varios árboles mediante voto mayoritario y los entrena con muestreo bootstrap (bagging), es el segundo modelo. Esta táctica disminuye la variabilidad del modelo y colabora en el control del ajuste excesivo al utilizar datos financieros históricos para la capacitación. En la práctica, esta programación se realiza en Python, normalmente en un notebook (.ipynb), usando bibliotecas de ciencia de datos y aprendizaje automático (por ejemplo, scikit-learn) para entrenar, hacer predicciones y evaluar el rendimiento con indicadores como la matriz de confusión y el AUC (Imani et al., 2025).

Bloque 1: Importación de librerías y configuración de visualización

```
# 1. Instalaciones y librerías

import pandas as pd
import numpy as np
```

En primer lugar, se importan las bibliotecas que apoyan el análisis, procesamiento y visualización de datos.

Numpy y *pandas* se utilizan para leer bases de datos, gestionar tablas (DataFrame) y llevar a cabo operaciones numéricas en variables financieras.

```
import matplotlib.pyplot as plt
import seaborn as sns
```

Se cargan *seaborn* y *matplotlib.pyplot* para crear gráficos, ya que posibilitan la creación de visualizaciones con un formato estandarizado y estadísticas para analizar comparaciones entre clases y distribuciones.

```
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.ensemble import RandomForestClassifier
```

El módulo de aprendizaje automático incluye *scikit-learn*, que permite importar *plot_tree* y *DecisionTreeClassifier* para el entrenamiento y la representación gráfica de árboles de decisión, así como *RandomForestClassifier* para entrenar un conjunto de árboles y fusionar predicciones.

```
from sklearn.metrics import confusion_matrix, roc_auc_score
from sklearn.preprocessing import StandardScaler
import warnings
warnings.filterwarnings('ignore')
```

La matriz de confusión se calcula con *confusion_matrix* y la capacidad de discriminación se estima a través del AUC con *roc_auc_score*, lo cual se utiliza para evaluar el rendimiento del clasificador. *StandardScaler* también se incorpore como transformador para estandarizar variables numéricas en caso de que sea necesario igualar las escalas durante el proceso de preparación de datos.

```
# Estilo profesional para gráficas
plt.rcParams.update({
    'font.size': 10,
    'axes.titlesize': 14,
    'axes.labelsize': 11,
    'legend.fontsize': 10,
    'figure.titlesize': 16
})
```

Primero se usa *plt. rcParams. update({. })* para configurar parámetros globales de Matplotlib. Aquí se fija el tamaño de letra base *(font. size = 10)*, el tamaño de títulos de ejes *(axes. titlesize = 14)*, etiquetas de ejes *(axes. labelsize = 11)*, texto de la leyenda *(legend. fontsize = 10)* y título general de la figura *(figure. titlesize = 16)*. Al ser global, estas opciones se aplican a todas las gráficas que se generen después, evitando ajustar fuente en cada figura.

```
# Colores corporativos
COLOR_LOW = '#2C3E50'   # Azul oscuro (bajo riesgo)
COLOR_HIGH = '#E74C3C'  # Rojo (alto riesgo)
```

```
COLOR_NEUTRAL = '#1ABC9C'  # Verde turquesa
(neutro/positivo)
```

Se procede a definir tres colores en formato hexadecimal. *COLOR_LOW* para indicar bajo riesgo, *COLOR_HIGH* para riesgo elevado, y *COLOR_NEUTRAL*, que engloba información y datos positivos.

```
sns.set_palette([COLOR_LOW, COLOR_HIGH])
```

Mediante *sns.set_palette([COLOR_LOW, COLOR_HIGH])* la paleta predefinida de *seaborn* se configura, y es por eso sus gráficos emplearán dichos colores de forma automática.

```
sns.set_style("whitegrid")
```

Para concluir, con *sns.set_style("whitegrid")* se implanta un estilo de fondo con una rejilla clara, muy útil al leer escalas, también para comparar valores en barras, líneas, o dispersión.

Bloque 2: Carga del archivo y creación del DataFrame

```
# 2. Carga y limpieza de datos
from google.colab import files
```

Dentro del entorno de ejecución, la carga inicial de la base de datos se lleva a cabo en este bloque. El comando *from google.colab import files* importa el módulo exclusivo de Google Colab que posibilita la administración de archivos desde el dispositivo del usuario.

```
uploaded = files.upload(
```

Después, *uploaded = files.upload()* abre un cuadro de selección para que el usuario pueda cargar el archivo en la zona temporal del cuaderno. El resultado se almacena en la variable *uploaded*, que funciona como un registro de los archivos que se han subido a lo largo de la sesión.

```
df = pd.read_excel("TESIS.xlsx", engine='openpyxl')
```

Por último, *df = pd.read_excel("TESIS.xlsx", enginer='openpyxl')* lee el archivo de Excel "TESIS.xlsx" con openpyxl como motor de lectura y lo transforma en un DataFrame denominado *df*, que será la estructura central para llevar a cabo procesos de depuración, transformación de variables y entrenamiento de los modelos de clasificación.

Bloque 3: Estandarización y limpieza de columnas numéricas con formato monetario

```
# Limpieza de columnas numéricas
def clean_currency(x):
```

Este bloque elabora columnas financieras que pueden incluir símbolos o separadores particulares de Excel con el fin de transformarlas a un formato numérico. La rutina de limpieza que se aplica a cada celda se denomina *clean_currency(x).*

```
if pd.isna(x):
        return np.nan
```

En primer lugar, *if pd.isna(x): return np.nan* reconoce valores ausentes (NaN) y los conserva como *np.nan* para prevenir fallos en las conversiones posteriores.

```
 if isinstance(x, str):
        return float(x.replace('$', '').replace(',', 
'').replace(' ', ''))
```

Luego, *if isinstance(x, str):* verifica si el dato está como texto; en ese caso, *x.replace('$', '').replace(',', '').replace(' ', '')* elimina el símbolo de dólar, las comas usadas como separador de millas y espacios en blanco, dejando solo dígitos y el punto decimal si existe.

```
    return float(x)
```

Después se convierte a *float(...)* para garantizar que el resultado sea numérico. La La función pasa directamente a *return float(x)* para garantizar un tipo de dato *float* uniforme en el caso de que el valor ya sea fuera numérico (no una cadena).

```
for col in ['Precio', 'Transporte', 'Seguro']:
```

El ciclo *for col in ['Precio', 'Transporte', 'Seguro']:* itera sobre las columnas elegidas que se deben manejar como montos.

```
df[col] = df[col].apply(clean_currency)
```

La función *clean_currency* se aplica a todos los registros de cada columna y la columna original es sustituida por su versión limpia, esto es: *df[col] = df[col].apply(clean_currency).*

El resultado es que *Precio*, *Transporte* y *Seguro* están listos para cálculos futuros, la creación de variables y el entrenamiento del modelo, sin ninguna inconsistencia debida al formato monetario o a las celdas introducidas como texto.

Bloque 4: Cálculo de variables derivadas y métricas por unidad

En este bloque se crean nuevas variables numéricas a partir de columnas existentes, con el fin de obtener indicadores comparables entre registros.

```
# Calcular métricas por unidad
df['precio_unitario'] = df['Precio'] / df['Cantidad']
```

La línea *df['precio_unitario'] = df['Precio'] / df['Cantidad']* calcula el precio por unidad, dividiendo el valor total de compra entre la cantidad adquirida; esta transformación permite comparar registros aunque tengan volúmenes distintos.

```
df['peso_unitario'] = df['Peso neto'] / df['Cantidad']
```

Luego, *df['peso_unitario'] = df['Peso neto'] / df['Cantidad']* estima el peso por unidad, útil cuando la cantidad se asocia a unidades físicas y se busca estandarizar el peso neto por cada unidad registrada.

```
df['costo_logistico_ratio'] = (df['Transporte'] + df['Seguro']) / df['Precio']
```

La línea *df['costo_logistico_ratio'] = (df['Transporte'] + df['Seguro']) / df['Precio']* construye un ratio de costo logístico, sumando los costos de transporte y seguro y dividiéndolos por el precio total.

El indicador expresa qué proporción del valor del producto corresponde a logística, en forma de fracción o porcentaje según se interprete, lo que facilita identificar operaciones con mayor carga logística relativa y usar esa información como variable explicativa en el modelo predictivo.

Bloque 5: Tratamiento de valores infinitos y estandarización de faltantes

Este bloque rectifica valores que no son válidos y que pueden surgir después de dividir variables, especialmente si *Cantidad* o *Precio* presentan ceros o datos ausentes.

```
# Tratar infinitos y NaN
df['precio_unitario'].replace([np.inf, -np.inf], np.nan, inplace=True)
df['peso_unitario'].replace([np.inf, -np.inf], np.nan, inplace=True)
df['costo_logistico_ratio'].replace([np.inf, -np.inf], np.nan, inplace=True)
```

Cada orden *replace([np.inf, -np.inf], np.nan, inplace=True)* busca los resultados infinitos positivos o negativos (*np.inf*, *-np.inf*) en la columna y los reemplaza con *np.nan*. Este último es el valor estándar para representar datos que faltan o no están definidos cuando se realiza análisis con pandas. El parámetro *inplace=True* realiza el reemplazo directamente en la columna, sin generar una copia extra.

Para que estas variables derivadas queden en un estado consistente antes de cualquier imputación, filtrado o entrenamiento del modelo, se repite el mismo procedimiento para *precio_unitario*, *peso_unitario* y *costo_logistico_ratio*. Esto previene que los algoritmos de aprendizaje automático y las métricas posteriores se equivoquen debido a valores infinitos, manteniendo el proceso de preparación de datos bajo control con una representación uniforme de situaciones problemáticas.

Bloque 6: Detección de valores extremos y construcción de la variable objetivo de riesgo

Este bloque determina límites estadísticos para distinguir valores extremos y posteriormente establece una etiqueta binaria de "alto riesgo".

```
p95_precio = df['precio_unitario'].quantile(0.95)
p05_precio = df['precio_unitario'].quantile(0.05)
```

Las líneas *p95_precio = df['precio_unitario'].quantile(0.95)* y *p05_precio = df['precio_unitario'].quantile(0.05)* determinan el percentil 95 y el 5 del precio unitario, respectivamente; estos valores son usados como límites de

referencia para señalar precios extraordinariamente bajos o altos en la distribución del conjunto de datos.

```
p90_logistico = df['costo_logistico_ratio'].quantile(0.90)
```

El percentil 90 del ratio logístico se obtiene con la línea *p90_logistico = df['costo_logistico_ratio'].quantile(0.90)*; este bloque lo deja calculado como un posible umbral. Sin embargo, la regla que se aplica más adelante utiliza un criterio fijo (0,25), lo que posibilita la comparación posterior entre un umbral "por regla" y uno "por percentil".

```
df['riesgo_alto'] = (
```

A continuación, se genera la columna *df['riesgo_alto']* a través de una condición compuesta. La expresión emplea operadores lógicos | (OR) para asignar un alto riesgo si se cumple alguna regla. Estas reglas son las siguientes:

```
    (df['costo_logistico_ratio'] > 0.25)
|                    # Costo logístico alto
```

costo_logístico_ratio > 0.25 indica operaciones donde el gasto de transporte y seguro es una fracción alta del precio;

```
    (df['precio_unitario'] > p95_precio) |                    #
Precio unitario muy alto
    (df['precio_unitario'] < p05_precio) |                    #
Precio unitario muy bajo (posible error)
```

precio_unitario > p95_precio y *precio_unitario < p05_precio* señalan extremos inferiores y superiores, en los que los valores bajos también pueden estar vinculados a errores de codificación o registro.

```
    (df['Proveedor'].isin(df['Proveedor'].value_counts().head(3
).index))  # Proveedores dominantes
```

Por último, *df['Proveedor'].isin(df['Proveedor'].value_counts().head(3).index)* señala registros que pertenecen a los tres proveedores con mayor frecuencia, estableciendo así un criterio de "concentración" por proveedor.

```
).astype(int)
```

El cierre *.astype(int)* transforma el resultado booleano (True/False) en 1 y 0, lo que permite utilizar una variable como variable objetivo en un clasificador.

```
print(f"☑ Total de transacciones: {len(df)}")
print(f"⚠ Transacciones con riesgo alto:
{df['riesgo_alto'].sum()} ({df['riesgo_alto'].mean():.1%})")
```

El total de transacciones y la cantidad/porcentaje que se clasifica como riesgo alto se resumen en las dos líneas *print*, lo cual es útil para revisar la proporción de clases antes de entrenar modelos.

Bloque 7: Selección de variables predictoras numéricas y categóricas

En esta sección se define el grupo de variables que nutrirán al modelo predictivo, organizándolas por tipo de dato para gestionarlas apropiadamente en el preprocesamiento.

```
# Seleccionar variables numéricas y categóricas
features_num = ['precio_unitario', 'peso_unitario',
'costo_logistico_ratio', 'Cantidad', 'Peso neto']
```

La lista *feature_num* incluye las variables numéricas: *costo_logistico_ratio* (una métrica derivada que estandariza el costo y el peso por unidad), *precio_unitario* y *peso_unitario*, además de las cantidades originales registradas, que son el *Peso neto* y la *Cantidad*. Esta La selección posibilita la captura de datos económicos relacionados con cada transacción, incluyendo la escala, el alcance y las proporciones.

```
features_cat = ['Pais de origen', 'Tipo de mercancia',
'Modelo']
```

Las variables categóricas se agrupan en la lista *features_cat*: *Modelo*, *Tipo de mercancía* y *País de origen*. Estas columnas no son cantidades, sino más bien clases o etiquetas; por lo tanto, se definen separadamente. Esto es así porque para modelos entrenar en scikit-learn generalmente se necesita convertirlas en variables numéricas a través de la codificación (como el one-hot encoding, por ejemplo).

Dividir datos numéricos y categóricos hace más fácil implementar transformaciones concretas a cada conjunto y conservar un flujo de

preparación de datos organizado antes del adiestramiento del árbol de decisión y del Random Forest.

Bloque 8: Construcción del dataset de modelado y depuración de la variable objetivo

En este bloque se arma una tabla específica para el entrenamiento del modelo, seleccionando únicamente las columnas necesarias.

```
# Codificación de variables categóricas
df_model = df[features_num + features_cat +
['riesgo_alto']].copy()
```

La instrucción *df_model = df[features_num + features_cat + ['riesgo_alto']].copy()* toma del DataFrame original *df* todas las variables predictoras numéricas *(features_num)*, las variables categóricas *(features_cat)* y la variable objetivo *riesgo_alto*. El método *.copy()* crea una copia independiente para trabajar el preprocesamiento sin alterar la base original, lo que facilita trazabilidad y evita cambios accidentales en *df.*

```
df_model = df_model.dropna(subset=['riesgo_alto'])
```

Luego, *df_model = df_model.dropna(subset=['riesgo_alto'])* elimina filas donde *riesgo_alto* esté vacío (NaN). Esta depuración se aplica solo sobre la etiqueta, ya que un modelo supervisado requiere que la variable objetivo esté definida en cada observación de entrenamiento.

El resultado es un conjunto de datos preparado para la etapa de codificación y transformación, con registros que cuentan con etiqueta válida para el aprendizaje.

Bloque 9: Codificación de variables categóricas y separación de predictores y variable objetivo

La información categórica se convierte a un formato numérico que es compatible con los algoritmos de clasificación en este bloque.

```
# Codificación categórica
df_encoded = pd.get_dummies(df_model, columns=features_cat,
drop_first=True)
```

La Instrucción *df_encoded = pd.get_dummies(df_model, columns=features_cat*, *drop_first=True)* implementa la codificación one-hot en

las columnas categóricas que están especificadas en *features_cat*. Este proceso genera nuevas columnas binarias (0/1) para cada categoría existente, lo que posibilita que el modelo reconozca país, tipo de mercancía y modelo como variables explicativas.

Al establecer el parámetro *drop_first=True*, se suprime la primera categoría de cada variable para prevenir la redundancia lineal entre las variables ficticias. Esto tiene como resultado que una categoría quede como referencia y se reduzca la colinealidad en el conjunto de predictores.

```
X = df_encoded.drop(columns=['riesgo_alto'])
y = df_encoded['riesgo_alto']
```

Después se divide el conjunto de datos en matriz de variables independientes y vector objetivo. Se crea *X* con todas las columnas que predicen, excepto la etiqueta, usando *X = df_encoded.drop(columns=['riesgo_alto'])*. La variable dependiente que el modelo intentará prever se obtiene con *y = df_encoded['riesgo_alto']*.

```
# Guardar nombres de columnas para la calculadora
X_columns = X.columns.tolist()
```

Por último, la línea de código *X_columns = X.columns.tolist()* almacena el listado de nombres de columnas obtenidas después de realizar la codificación. Se utiliza para asegurar la coherencia cuando se aplica el modelo a datos nuevos, para elaborar informes sobre la relevancia de las variables o para diseñar una calculadora en la que las entradas deben tener el mismo orden y las mismas columnas que se emplearon durante el entrenamiento.

Bloque 10: Entrenamiento y predicción con Random Forest

En este bloque se determina y se entrena el modelo Random Forest con el propósito de categorizar el riesgo alto (*riesgo_alto*) basándose en las variables predictoras *X*.

```
# Random Forest
rf = RandomForestClassifier(n_estimators=100,
random_state=42, class_weight='balanced',max_depth=3)
```

La línea que establece el objeto del modelo, *rf = RandomForestClassifier(n_estimators=100, random_state=42, class_weight='balanced', max_ Depth=3)*, utiliza parámetros específicos:

n_estimators=100 señala que 100 árboles serán entrenamientos, lo cual posibilita la media de los resultados y la estabilización de la predicción; *random_state=42* define una semilla aleatoria para garantizar que el entrenamiento sea reproducible; *class_weight='balanced'* modifica el peso de las clases según su frecuencia, otorgando mayor peso a la clase menos frecuente con el fin de disminuir sesgos cuando hay escasos casos de "alto riesgo"; *max_depth=3* acota la profundidad de cada árbol con el objetivo de prevenir sobreajuste y sostener reglas más sencillas.

```
rf.fit(X, y)
```

Después, *rf.fit(X, y)* entrena el bosque con la matriz de predictores *X* junto a la etiqueta *y*. En este proceso, cada árbol se construye a partir de una muestra aleatoria de los datos (bootstrap) y con la elección aleatoria de variables en las divisiones, lo cual produce diversidad entre los árboles.

```
rf_pred = rf.predict(X)
```

Por último, *rf_pred = rf.predict(X)* logra las predicciones de clase (0 o 1) para los datos que se emplearon en el entrenamiento. Este paso permite realizar una primera evaluación del comportamiento del modelo; sin embargo, para examinar el rendimiento real, se sugiere predecir a partir de un conjunto de prueba separado o usando validación cruzada, puesto que las predicciones sobre *X* suelen dar lugar a un desempeño optimista.

Bloque 11: Entrenamiento y predicción con Árbol de Decisión

Un Árbol de Decisión es desarrollado en este bloque para clasificar la variable *riesgo_alto* calculando en los predictores *X*.

```
# Árbol de decisión
dt = DecisionTreeClassifier(max_depth=3, random_state=42,
class_weight='balanced')
```

La línea *dt = DecisionTreeClassifier(max_ Depth=3, random_state=42, class_weight='balanced')* instancia el modelo con parámetros que buscan control y reproducibilidad. *max_ Depth=3* establece un límite a la profundidad del árbol, lo que reduce el número de divisiones posibles y previene que el modelo memorice el conjunto de entrenamiento. Esto mejora la interpretabilidad, ya que las reglas del árbol son más breves.

La semilla del proceso de particionamiento queda fijada por *random_state=42*, lo que permite obtener resultados que se pueden repetir. *class_weight='balanced'* mitiga la falta de equilibrio en las clases al otorgar pesos inversamente proporcionales a su frecuencia, lo que hace que el árbol preste más atención a la clase con menos representación.

```
dt.fit(X, y)
```

Después, *dt.fit(X, y)* entrena el árbol con los datos codificados; el algoritmo busca divisiones en las variables que disminuyen la impureza (como Gini, por ejemplo) y logran una separación más efectiva de las clases.

```
dt_pred = dt.predict(X)
```

Finalmente, *dt_pred = dt.predict(X)* produce predicciones de clase (0 o 1) sobre el mismo conjunto *X*. Este resultado se utiliza a menudo para verificar de manera preliminar cómo funciona el modelo; Sin embargo, la evaluación técnica se lleva a cabo con datos de prueba o validación cruzada, pues la Práctica en el conjunto de entrenamiento no muestra qué tan bien va el modelo fuera de la muestra.

Bloque 12: Cálculo e impresión de métricas de desempeño del modelo

Este bloque presenta un resumen de ejecución para los dos modelos, utilizando el mismo conjunto de entrenamiento.

```
print("\n📊 Métricas (entrenamiento completo):")
```

La única función de la línea *print("\n📊 Métricas (entrenamiento completo):")* es mostrar un encabezado en pantalla. Después, se compara la etiqueta real con la predicción de clase (0/1) de cada modelo y se calcula el AUC mediante *roc_auc_score(y, rf_pred)* y *roc_auc_score(y, dt_pred)*.

```
print(f"Random Forest AUC: {roc_auc_score(y,
rf_pred):.3f}")
print(f"Árbol de Decisión AUC: {roc_auc_score(y,
dt_pred):.3f}")
print(f"Random Forest Accuracy: {rf.score(X, y):.3f}")
print(f"Árbol de Decisión Accuracy: {dt.score(X, y):.3f}")
print(f"Random Forest Precision: {rf.score(X, y):.3f}")
print(f"Árbol de Decisión Precision: {dt.score(X, y):.3f}")
print(f"Random Forest Recall: {rf.score(X, y):.3f}")
print(f"Árbol de Decisión Recall: {dt.score(X, y):.3f}")
```

```
print(f"Random Forest F1: {rf.score(X, y):.3f}")
print(f"Árbol de Decisión F1: {dt.score(X, y):.3f}")
```

Para mostrar tres decimales, el valor se formatea con la instrucción:*.3f.* A pesar de que el AUC mide la habilidad para distinguir clases en términos de interpretación, su cálculo estándar se realiza con probabilidades (*predict_proba*) o calificación, no con clase dura 0/1. Por esta razón, es posible que el AUC aquí sea menos informativo de lo esperado y sea recomendable modificarlo en la versión final.

Se imprime "Accuracy" a partir de *rf.score(X, y)* y *dt.score(X, y),* que en scikit-learn se refiere a la proporción de aciertos sobre *X*. En esta secuencia de código, las líneas marcadas como Precision, Recall y F1 vuelven a emplear *score(X, y),* así que no están calculando realmente precisión, recuerda ni F1; simplemente están repitiendo la exactitud con diferentes etiquetas.

Para reportar esas métricas de manera adecuada, es necesario emplear funciones concretas como *f1_score*, *precision_score* y *recall_score* junto con *y* y la predicción *(rf_pred, dt_pred).* Dado que los valores tienden a ser artificialmente altos al evaluarlos sobre el entrenamiento, también es recomendable calcular estas métricas mediante validación cruzada o en un conjunto de prueba.

Bloque 13: Visualización de la distribución de transacciones por nivel de riesgo

Este bloque crea un gráfico de barras que resume la cantidad de transacciones clasificadas como alto riesgo y bajo riesgo, calculando en la variable *riesgo_alto*.

```
fig, ax = plt.subplots(figsize=(8, 5))
```

La instrucción *fig, ax = plt.subplots(figsize=(8, 5))* genera el eje y la figura con dimensiones preestablecidas para su presentación.

```
counts = df['riesgo_alto'].value_counts()
```

Después, *counts = df['riesgo_alto'].value_counts()* determina cuántas veces aparece cada clase (1 y 0).

```
bars = ax.bar(['Bajo riesgo', 'Alto riesgo'], counts,
color=[COLOR_LOW, COLOR_HIGH], edgecolor='white', linewidth=1.2)
```

Se crea dos barras con *ax.bar(['Bajo riesgo', 'Alto riesgo'], counts, color=[COLOR_LOW, COLOR_HIGH], edgecolor='white', linewidth=1.2)*, utilizando colores que han sido establecidos antes y borde blanco para lograr un acabado más pulido en el informe.

```
    ax.set_title('Distribución de Transacciones por Nivel de
Riesgo', fontsize=14, pad=20)
    ax.set_ylabel('Número de transacciones')
    ax.set_ylim(0, max(counts) * 1.1)
```

Luego, se modifica la lectura del gráfico: *ax.set_title(...)* sitúa el título con un espacio superior *(pad=20)*, *ax.set_ylabel(...)* asigna nombre al eje vertical, y *ax.set_ylim(0, max(counts) * 1.1)* aumenta el límite superior para proporcionar espacio a las etiquetas numéricas.

```
    for bar in bars:
        height = bar.get_height()
        ax.annotate(f'{int(height):,}', xy=(bar.get_x() +
bar.get_width() / 2, height),
                    xytext=(0, 3), textcoords="offset points",
ha='center', va='bottom', fontweight='bold')
```

El ciclo *for bar in bars:* examina cada barra y obtiene su altura utilizando *bar.get_height()*; después, *ax.annotate(...)* escribe el valor de cada barra en la parte superior de ella con formato de miles *({int(height):,})* y tipografía en negrita.

```
    sns.despine(left=True, bottom=True)
    plt.tight_layout()
    plt.show()
```

Por último, *sns.despine(left=True, bottom=True)* suprime los bordes del gráfico para lograr una apariencia más ordenada; *plt.tight_layout()* modifica los márgenes para prevenir que el texto se corte; y *plt.show()* muestra el resultado en la pantalla.

Bloque 14: Cálculo y visualización de la importancia de variables en Random Forest

Este bloque muestra en un gráfico las variables que más contribuyen a las decisiones del modelo Random Forest.

```
    # Importancia de variables
    importances = rf.feature_importances_
```

La línea *importances = rf.feature_importances_* obtiene el atributo del modelo, ya entrenado, que incluye la importancia relativa de cada predictor. En Random Forest, este valor se calcula en función de cuánto disminuye la impureza cada variable a través de los árboles.

```
indices = np.argsort(importances)[::-1][:12]
top_features = [X.columns[i] for i in indices]
top_importances = importances[indices]
```

Después, las importancias se ordenan de mayor a menor con el *comando indices=np.argsort(importances)[::-1][:12]*: *np.argsort* devuelve los índices que ordenarían el arreglo; *[::-1]* lo invierte para que esté en orden decreciente; *[:12]* Escoge únicamente las 12 variables más importantes. Con *top_features = [X.columns[i] for i in indices]*, se obtienen los nombres de las columnas en el mismo orden, y con *top_importances = imports[indices]*, se almacenan los valores numéricos correspondientes para ser graficados.

```
fig, ax = plt.subplots(figsize=(10, 6))
```

En la parte gráfica, *fig, ax = plt.subplots(figsize=(10, 6))* genera una figura grande que permite el ajuste de etiquetas largas.

```
bars = ax.barh(range(len(top_importances)),
top_importances, color=COLOR_NEUTRAL, edgecolor='white',
linewidth=0.8)
```

La instrucción *ax.barh(...)* crea barras horizontales, lo cual es útil para facilitar la lectura si hay muchos nombres. Se emplean bordes blancos y *COLOR_NEUTRAL* para mantener un estilo uniforme.

```
ax.set_yticks(range(len(top_importances)))
ax.set_yticklabels(top_features)
```

Luego, con el método *ax.set_yticks(...)* se ajustan las etiquetas del eje Y, utilizando también *ax.set_yticklabels(top_features*).

```
ax.invert_yaxis()
```

Por último, se invierte el orden con *ax.invert_yaxis()* para que la variable más relevante quede en la parte superior.

```
ax.set_xlabel('Importancia relativa')
ax.set_title('Top 12 Variables que Impulsan el Riesgo
Operativo', fontsize=14, pad=20)
```

Se designan el eje y el gráfico a través de las funciones *ax.set_xlabel('Importancia relativa')* y *ax.set_title(...).*

```
    or i, v in enumerate(top_importances):
        ax.text(v + 0.001, i, f"{v:.3f}", va='center',
fontweight='bold')
```

El bucle *for i, v in enumerate(top_importances): ax.text(...)* pone el valor numérico de cada barra a la derecha, con tres decimales, lo que hace que sea más fácil de interpretar.

```
    sns.despine(left=True, bottom=True)
    plt.tight_layout()
    plt.show()
```

Al final, *sns.despine(...)* limpia los bordes, *plt.tight_layout()* adapta los márgenes y *plt.show()* presenta el gráfico.

Bloque 15: Matriz de confusión comparativa para Random Forest y Árbol de decisión

```
    fig, axes = plt.subplots(1, 2, figsize=(14, 5))
```

Esta sección genera una representación visual comparativa de la actuación de los dos modelos, utilizando matrices de confusión que se presentan como mapas térmicos. La figura con dos subgráficos en una fila, uno para el modelo de Árbol de Decisión y otro para el modelo de Random Forest, es generado por la instrucción *fig, axes = plt.subplots(1, 2, figsize=(14, 5))*.

```
    for idx, (pred, name, color) in enumerate(zip([rf_pred,
dt_pred], ['Random Forest', 'Árbol de Decisión'], [COLOR_NEUTRAL,
COLOR_LOW])):
```

Después, se emplea un ciclo *for idx, (pred, name, color) in enumerate(zip([...])):* que itera simultáneamente sobre las predicciones (*rf_pred* y *dt_pred*) y los nombres de cada modelo.

```
    cm = confusion_matrix(y, pred)
```

En cada repetición, se computa la matriz de confusión a través de *cm = confusion_matrix(y, pred)*, lo que da como resultado una tabla 2×2 que resume los errores y logros del clasificador: verdaderos negativos, verdaderos positivos, falsos negativos y falsos positivos, en el orden estándar de scikit-learn.

```
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
cbar=False, ax=axes[idx],
                annot_kws={"size": 12, "weight": "bold"})
```

A continuación, se dibuja la matriz con *sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False, ax=axes[idx], ...)*, mostrando los conteos dentro de cada celda en la que se exhiben los conteos de cada celda (*annot=True*)como enteros (*fmt='d'*) y utilizando una paleta azul sin barra lateral de color para preservar el orden visual.

```
    axes[idx].set_title(f'{name}\nAUC: {roc_auc_score(y,
pred):.3f}', fontsize=13, pad=15)
        axes[idx].set_xlabel('Predicción')
        axes[idx].set_ylabel('Real')
```

En cada panel se configuran títulos y ejes: *axes[idx].set_title(...)* agrega el nombre del modelo y un valor AUC determinado con *roc_auc_score(y, pred).* En cambio, *set_xlabel('Predicción')* y *set_ylabel('Real')* señalan lo que representa cada eje.

```
    plt.suptitle('Desempeño de los Modelos en la Detección de
Riesgo', fontsize=15, y=1.02)
    plt.tight_layout()
    plt.show()
```

Por último, *plt.suptitle(...)* añade un título general a la figura entera, *plt.tight_layout()* configura automáticamente los espacios para que los textos no se superpongan y *plt.show()* muestra el resultado.

Bloque 16: Visualización e interpretación de reglas del Árbol de Decisión

Se produce una representación gráfica del árbol de decisión que se ha entrenado para entender sus normas de clasificación.

```
    # Árbol de decisión
    plt.figure(figsize=(22, 12))
```

La figura que se genera con la instrucción *plt.figure(figsize=(22, 12))* es lo suficientemente grande como para que el diagrama del árbol sea legible en caso de que existan múltiples nodos y textos.

```
    plot_tree(dt, feature_names=X.columns, class_names=['Bajo
Riesgo', 'Alto Riesgo'], max_depth=5,
```

Después, *plot_tree(dt, feature_names=X.columns, class_names=['Bajo Riesgo', 'Alto Riesgo'], max_depth=5, filled=True, rounded=True, fontsize=10, proportion=True, impurity=False)* representa gráficamente la estructura del

modelo *dt.* Se emplea *feature_names=X.columns* para revelar el nombre verdadero de cada variable en las particiones.

Se pueden asignar etiquetas que se puedan interpretar a las clases usando *class_names=['Bajo Riesgo', 'Alto Riesgo'].* El modelo entrenado no se altera, pero la visualización se limita a cinco niveles de profundidad para preservar la legibilidad.

```
 filled=True, rounded=True, fontsize=10,
          proportion=True, impurity=False)
```

En la información y en el aspecto visual, *fill=True* colorea los nodos de acuerdo con la clase mayoritaria, *rounded=True* redondea los bordes para que la presentación sea mejor y *fontsize=10* regula el tamaño del texto que está dentro de cada nodo. Cuando *proportion=True*, las cifras que aparecen en los nodos se expresan como proporciones del total de cada nodo, lo cual facilita la interpretación de la composición de clases en cada regla.

```
plt.title("Reglas de Negocio para Identificar Riesgo
Operativo\n(Árbol de Decisión Interpretado)",
          fontsize=16, pad=30, fontweight='bold')
plt.tight_layout()
plt.show()
```

Para simplificar el diagrama y priorizar la lectura de reglas, se oculta la impureza (ya sea entropía o Gini) cuando *impurity=False*. Finalmente, *plt.title(...)* pone un título descriptivo con salto de línea; *plt.tight_layout()* modifica los márgenes para evitar que se recorten textos y *plt.show()* muestra el gráfico. Este esquema se emplea como fundamento para escribir "reglas de negocio" del estilo: si determinadas variables exceden o no un límite, la transacción se clasifica, en consecuencia, como de riesgo alto o bajo.

Bloque 17: Ranking de riesgo por proveedor con base en predicción del Árbol de Decisión

Este bloque elabora un ranking de proveedores de acuerdo con lo que el árbol de decisión estima como riesgo.

```
#8. Ranking de riesgo por proveedor
df['pred_dt'] = dt.predict(X)
```

Para cada registro de *X*, el código *df['pred_dt'] = dt.predict(X)* produce pronósticos (0 = bajo riesgo, 1 = alto riesgo), que se almacenan en una nueva

columna del DataFrame original *df*, denominada *pred_dt*. De esta manera, el modelo de etiqueta de cada transacción, lo que posibilita un análisis global por proveedor.

```
df_proveedor = df.groupby('Proveedor').agg(
    total_transacciones=('pred_dt', 'count'),
    transacciones_riesgo=('pred_dt', 'sum'),
    gasto_total=('Precio', 'sum')
).reset_index()
```

A continuación, *df_proveedor = df.groupby('Proveedor').agg(...)* reúne todas las transacciones de acuerdo con el campo *Proveedor* y determina tres indicadores: *total_transacciones* computa cuántos registros posee cada proveedor *(count)*; *transacciones_riesgo* agrega las predicciones *(sum)*, que corresponde al total de transacciones categorizadas como riesgo alto; y *gasto_total* suma la columna *Precio* para calcular el volumen económico vinculado a dicho proveedor. La instrucción. reset_*index()* transforma el resultado del agrupamiento en una tabla normal, manteniendo la columna de *Proveedor.*

```
df_proveedor['tasa_riesgo'] =
df_proveedor['transacciones_riesgo'] /
df_proveedor['total_transacciones']
```

Luego se calcula una tasa que se puede comparar entre proveedores: *df_proveedor['tasa_riesgo'] = ...* divide las operaciones de alto riesgo entre el total de transacciones, generando un valor que oscila entre 0 y 1. Esto puede entenderse como la proporción de operaciones catalogadas como de alto riesgo para cada proveedor.

```
df_proveedor = df_proveedor.sort_values('tasa_riesgo',
ascending=False)
```

Después, *df_proveedor = df_proveedor.sort_values('tasa_riesgo', ascending=False)* organiza la tasa de mayor a menor para que los proveedores con una mayor concentración de riesgo aparezcan primero.

```
print("\n Proveedores con mayor tasa de riesgo:")
print(df_proveedor[['Proveedor', 'total_transacciones',
'tasa_riesgo', 'gasto_total']].head(10))
```

Por último, los *print* exhiben la tabla resumida y el encabezado con las primeras 10 filas *(head(10))*, que contienen el total de transacciones, el

proveedor, la tasa de riesgo y el gasto total. Esto permite darle prioridad a la revisión por proporción de riesgo, frecuencia y monto asociado.

Bloque 18: Cálculo de KPI de riesgo y clasificación de severidad

Este bloque crea un indicador de resumen (KPI) basado en el porcentaje de transacciones clasificadas como de alto riesgo.

```
total_transacciones = len(df)
alto_riesgo = df['riesgo_alto'].sum()
```

El total de registros en el DataFrame se determina con la línea *total_transacciones = len(df)*. Después, *alto_riesgo = df['riesgo_alto'].sum()* suma la variable binaria riesgo_alto, que tiene un valor de 1 si el riesgo es alto. Esta suma da como resultado la cantidad de transacciones marcadas como de alto riesgo.

```
kpi_riesgo = alto_riesgo / total_transacciones
```

El KPI se calcula como una tasa entre 0 y 1, utilizando la fórmula *kpi_riesgo = alto_riesgo / total_transacciones.* Esta tasa puede interpretarse como el porcentaje de operaciones de alto riesgo respecto al total.

```
print(f"📊 KPI de Riesgo Operativo: {kpi_riesgo:.2%}")
print(f"   - Transacciones de alto riesgo:
{alto_riesgo:,}")
print(f"   - Total de transacciones:
{total_transacciones:,}")
```

Las instrucciones de *print* muestran el KPI en porcentaje *(:.2%)* y reportan las cifras totales y las de alto riesgo, utilizando un separador de millas para facilitar la lectura.

```
# Clasificación de severidad
if kpi_riesgo < 0.10:
    nivel = "🟢 Bajo"
elif kpi_riesgo <= 0.25:
    nivel = "🟡 Moderado"
else:
    nivel = "🔴 Alto"

print(f"   - Nivel de severidad: {nivel}")
```

La gravedad del KPI se clasifica en la segunda parte usando reglas por umbrales. Según el valor de *kpi_riesgo*, el bloque *if/elif/else* asigna un nivel cualitativo: "Bajo" si es menor a 0.10, "Moderado"si está entre 0,10 y 0,25, y

"Alto" si es mayor a 0,25. Esta clasificación ayuda a priorizar y a interpretar desde el punto de vista gerencial, dado que convierte un porcentaje en un semáforo de riesgo para su reporte. La línea final imprime el nivel asignado con un ícono de color correspondiente, lo que permite una lectura rápida del resultado.

Bloque 19: Función de cálculo de riesgo operativo para nuevas transacciones

```
    def calcular_riesgo(precio, transporte, seguro, cantidad,
peso_neto, pais="CN", tipo="ACCESORIO", modelo="SEDÁN"):
        """
        Calcula el riesgo operativo de una transacción dada.

        Parámetros:
        - precio: Precio total (USD)
        - transporte: Costo de transporte (USD)
        - seguro: Costo de seguro (USD)
        - cantidad: Número de unidades
        - peso_neto: Peso total (kg)
        - pais: País de origen (ej. 'CN', 'JP')
        - tipo: Tipo de mercancía (ej. 'ACCESORIO', 'REPUESTO')
        - modelo: Modelo de vehículo (ej. 'SEDÁN', 'SUV')

        Retorna:
        - Predicción: 'Alto Riesgo' o 'Bajo Riesgo'
        - Probabilidad de alto riesgo
        """
        # Validaciones
        if cantidad <= 0 or precio <= 0:
            raise ValueError("La cantidad y el precio deben ser
mayores que cero.")
```

Este bloque establece una "calculadora" que facilita el cálculo del riesgo de una nueva transacción mediante el uso del modelo previamente entrenado, que en este caso es un Árbol de Decisión guardado en *dt.* La función *calcular_riesgo(...)* toma como entradas valores por defecto para simplificar las pruebas cuando no se introducen todos los parámetros, incluyendo montos (*precio*, *transporte*, *seguro*), cantidades físicas (*peso_neto*, *cantidad*) y campos categóricos (*tipo*, *país*, *modelo*).

El docstring explica cuál es la finalidad de la función, sus parámetros y lo que esta devuelve. Se realiza una validación básica antes de calcular: en

caso de que el *precio* o la *cantidad* sean cero o inferiores a este, se producirá un error (*ValueError)*, pues estas variables se aplicarán en divisiones y un cero produciría resultados indefinidos o infinitos.

```
# Calcular variables derivadas
    precio_unitario = precio / cantidad
    peso_unitario = peso_neto / cantidad
    costo_logistico_ratio = (transporte + seguro) / precio
```

Después, la función genera las mismas variables derivadas que se emplearon en el entrenamiento: *costo_logistico_ratio*, *precio_unitario* y *peso_unitario*. Esto garantiza la consistencia entre los nuevos datos y el patrón que el modelo ha aprendido.

```
# Crear DataFrame temporal
    temp_df = pd.DataFrame({
        'precio_unitario': [precio_unitario],
        'peso_unitario': [peso_unitario],
        'costo_logistico_ratio': [costo_logistico_ratio],
        'Cantidad': [cantidad],
        'Peso neto': [peso_neto],
        'Pais de origen': [pais],
        'Tipo de mercancia': [tipo],
        'Modelo de vehiculo': [modelo]
  })
```

Se genera un DataFrame temporal *temp_df* con una única fila a partir de esas variables; esta fila simboliza la transacción que se va a analizar y conserva los mismos nombres de las columnas que aparecieron en el conjunto de datos para el modelado.

```
 # Codificar categóricas
    temp_encoded = pd.get_dummies(
        temp_df,
        columns=['Pais de origen', 'Tipo de mercancia',
'Modelo de vehiculo'],
        drop_first=True
    )
```

Luego, se usa *pd.get_dummies(...)* para convertir las variables categóricas a formato binario, al igual que en la etapa de entrenamiento.

```
 # Asegurar alineación con las columnas del modelo
    for col in X_columns:
        if col not in temp_encoded.columns:
            temp_encoded[col] = 0
    temp_encoded = temp_encoded[X_columns]
```

Dado que al codificar una única fila algunas categorías pueden no estar presentes, aunque sí estaban en el conjunto de entrenamiento, se realiza un ciclo que recorre *X_columns* (el listado de columnas del modelo de entrenamiento) y agrega las columnas faltantes con ceros. De esta manera, se asegura que *temp_encoded* mantenga las mismas variables y el mismo orden que prevé el modelo.

```
        # Predecir directamente (sin escalado)
            pred = dt.predict(temp_encoded)[0]
            prob = dt.predict_proba(temp_encoded)[0, 1]

            return ("Alto Riesgo" if pred == 1 else "Bajo Riesgo",
prob)
```

Al final, *dt.predict(temp_encoded)[0]* proporciona la clase estimada (de 0 o 1), mientras que *dt.predict_proba(temp_encoded)[0, 1]* brinda la probabilidad vinculada a la clase "alto riesgo". La función devuelve una etiqueta comprensible ("Alto Riesgo" o "Bajo Riesgo"), además de la probabilidad, lo que posibilita su utilización como instrumento de evaluación rápida para transacciones nuevas sin necesidad de reentrenar el modelo.

Bloque 20: Ejemplo de ejecución de la calculadora y control de errores

```
#Calculadora

print("\n" + "="*60)
print("EJEMPLO DE USO DE LA CALCULADORA")
print("="*60)
```

El bloque muestra un ejemplo práctico de cómo se aplica la función *calcular_riesgo* y presenta el resultado en la consola con formato de informe. Las líneas iniciales crean separadores visuales con la instrucción *"="*60*, que produce una línea de 60 caracteres igual a "=", con el fin de marcar los límites de la sección y mejorar la legibilidad del resultado. Para ejecute la calculadora de manera segura, utilizando un bloque *try:*, se capturarán posibles errores de validación o datos sin que el cuaderno entero se detenga.

```
try:
    resultado, prob = calcular_riesgo(
        precio=10000, transporte=1000, seguro=20,
cantidad=100, peso_neto=50,
        pais="CN", tipo="REPUESTO", modelo="SEDÁN"
```

```
        )
        print(f"ANHUI AGGEUS AUTO -TECH CO., LTD: {resultado}
(prob: {prob:.2%})")
```

En el bloque *try*, se invoca *calcular_riesgo(...)* con ejemplos de valores para una operación: precio total, transporte, seguro, cantidad y peso neto, así como las categorías de país, tipo y modelo. La función proporciona dos salidas: una, la probabilidad calculada de alto riesgo *(prob)*; la otra, el *resultado*, que puede ser "Alto Riesgo" o "Bajo Riesgo". A continuación, se genera una línea de informe: *print(f"... {resultado} (prob: {prob:.2%})")*, donde la probabilidad se presenta como un porcentaje con dos decimales *(:.2%)* y está ligada al nombre del proveedor que aparece en el texto.

```
except Exception as e:
    print(f"❌ Error: {e}")
```

El bloque *except Exception as e:* atrapa cualquier excepción que ocurra dentro del *try* (como precio o cantidad no válidas, por ejemplo) y enseña un mensaje con el error, lo que posibilita la depuración de la entrada sin detener el flujo del análisis.

RESULTADOS

Análisis exploratorio de datos

Comenzando el análisis exploratorio, primero se muestra una descripción estructural del conjunto de datos utilizado para el modelado. En esta parte se puede identificar el tipo de información que se tiene, ya sea numérica o categórica, y la función que desempeña cada variable en el proceso (dato original, variable derivada o variable objetivo). Este paso permite hacer un seguimiento de las transformaciones posteriores y definir qué campos son los que directamente nutren a los modelos de clasificación.

Descripción de variables del dataset

Se elabora un conjunto de datos a partir de transacciones históricas de importación 2021-2025, que combinan información económica (precio, flete, seguro), operativa (cantidad, peso neto) y de identificación (país de origen, tipo de mercancía, modelo, proveedor). De Estas columnas se derivan métricas por unidad y ratios de coste que normalizan las comparaciones entre transacciones de diferente tamaño, para medir el riesgo en términos comparables entre operaciones y no en cifras absolutas. En la Tabla 1 se muestran las variables de acuerdo con el dataset.

Tabla 2
Descripción del dataset y variables

Variable	Tipo	Descripción
Precio	num	Monto neto de la compra (USD)
Transporte	num	Gasto de transporte por ítem (USD)
Seguro	num	Gasto de seguro por ítem (USD)
Cantidad	num	Unidades físicas en la transacción
Peso neto	num	Peso neto total de la transacción (kg)
Pais de origen	cat	Código/país de procedencia
Tipo de mercancia	cat	Categoría del bien importado
Modelo	cat	Segmento/modelo asociado
Proveedor	cat	Razón social del proveedor
precio_unitario	deriv	Precio / Cantidad (USD por unidad)
peso_unitario	deriv	Peso neto / Cantidad (kg por unidad)
costo_logistico_ratio	deriv	(Transporte + Seguro) / Precio
riesgo_alto	target	Etiqueta binaria construida (1=alto, 0=bajo)

La Tabla 1 muestra una estructura apropiada para la clasificación supervisada, ya que mezcla variables categóricas de importancia para

segmentar con variables cuantitativas. Las variables precio_unitario, peso_unitario y costo_logistico_ratio se convierten y cantidades físicas en indicadores normalizados, lo cual disminuye la dependencia de la escala y facilita la interpretación .de las compras pequeñas en comparación con las compras grandes.

La etiqueta binaria riesgo_alto establece la meta de aprendizaje, mientras que pred_dt permite comparar la predicción del árbol con dicha etiqueta; esto es ventajoso para verificar la consistencia a nivel de registro y realizar análisis posteriores en caso de error de clasificación.

Medidas descriptivas de variables numéricas

Para el análisis exploratorio se obtienen estadísticos descriptivos de las variables numéricas para caracterizar la distribución de los datos antes del modelado. La Tabla 2 resume los valores de media, mediana, moda, desviación estándar, varianza, mínimos, máximos, rangos y los distintos cuartiles obteniendo finalmente un valor de IQR los cuales se muestran a continuación.

Figura 5

Medidas descriptivas de variables numéricas

	Media	Mediana	Moda	Desviación estándar	Varianza	Mínimo	Máximo	Rango	Q1 (25%)	Q3 (75%)	IQR
Peso neto	622.530021	339.050000	160.000000	1036.959985	1.075286e+06	0.140000	17000.000000	16999.860000	102.340000	744.400000	642.060000
Año de fabricacion	2022.933789	2023.000000	2023.000000	1.382813	1.912171e+00	2021.000000	2025.000000	4.000000	2022.000000	2024.000000	2.000000
Cantidad	144.174365	68.000000	100.000000	486.733334	2.369093e+05	0.450000	10000.000000	9999.550000	30.000000	120.000000	90.000000
Precio	2227.879139	1358.000000	600.000000	3145.190626	9.892224e+06	0.100000	51743.000000	51742.900000	600.000000	2889.800000	2289.800000
Transporte	247.626912	99.880000	29.080000	408.722405	1.670540e+05	0.040000	4695.310000	4695.270000	32.345000	297.210000	264.865000
Seguro	12.564803	7.340000	6.000000	20.469010	4.189804e+02	0.001000	545.130000	545.129000	3.070000	15.275000	12.205000
precio_unitario	53.967293	28.500000	12.000000	472.273229	2.230420e+05	0.001429	23600.000000	23599.998571	9.000000	52.440000	43.440000
peso_unitario	14.527381	7.233333	1.000000	248.621405	6.181260e+04	0.001000	16350.000000	16349.999000	1.375600	11.950000	10.574400
costo_logistico_ratio	0.112045	0.081421	0.048858	0.086386	7.462490e-03	0.013005	0.515751	0.502746	0.048934	0.144192	0.095258
riesgo_alto	0.487751	0.000000	0.000000	0.499905	2.499051e-01	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000
pred_dt	0.396822	0.000000	0.000000	0.489292	2.394071e-01	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000

El resumen de la Figura 5 muestra las tendencias centrales y la dispersión observadas en las variables numéricas clave del conjunto de datos. Tanto en Precio (media 2.227,88; mediana 1.358; máximo 51.743) como en Cantidad (media 144,17; mediana 68; máximo 10.000) la discrepancia entre media y mediana junto a amplios rangos indican distribuciones asimétricas, salpicadas por transacciones de considerable magnitud. Similar patrón se repite en Peso neto (media 622,53; mediana 339,05; máximo 17.000) evidenciando alta dispersión (desviación estándar 1.036,96) y, además, un IQR (642,06) que confirma una notable variabilidad entre las transacciones. En los costos asociados Transporte (media 247,63; mediana 99,88; máximo 4.695,31) y Seguro (media 12,56; mediana 7,34; máximo 545,13), se aprecian también colas prolongadas, un fenómeno que coincide con las variaciones en rutas, volumen y condiciones logísticas.

En lo concerniente a las métricas derivadas, precio_unitario (mediana 28,5; Q1 9,0; Q3 52,44; máximo 23.600) y peso_unitario (mediana 7,23; Q1 1,38; Q3 11,95; máximo 16.350) dejan ver valores extremos notables capaces de influir los promedios; por eso, los cuartiles y el IQR son bastante útiles al describir el "rango habitual" sin verse abrumados por los valores atípicos. costo_logistico_ratio exhibe una mediana de 0,081 y un Q3 de 0,144 con un máximo de 0,516, un rango que se corresponde con las operaciones donde el transporte y el seguro conforman una fracción importante del precio final.

Finalmente, el riesgo_alto exhibe una media de 0,487751, que se traduce en un 48,78% de operaciones con alto riesgo. Consideremos además, pred_dt con una media de 0,396822, revelando que el árbol pronostica un alto riesgo en una fracción inferior a la etiqueta real. Este hallazgo, es bastante consistente con la naturaleza conservadora del modelo al activar alertas.

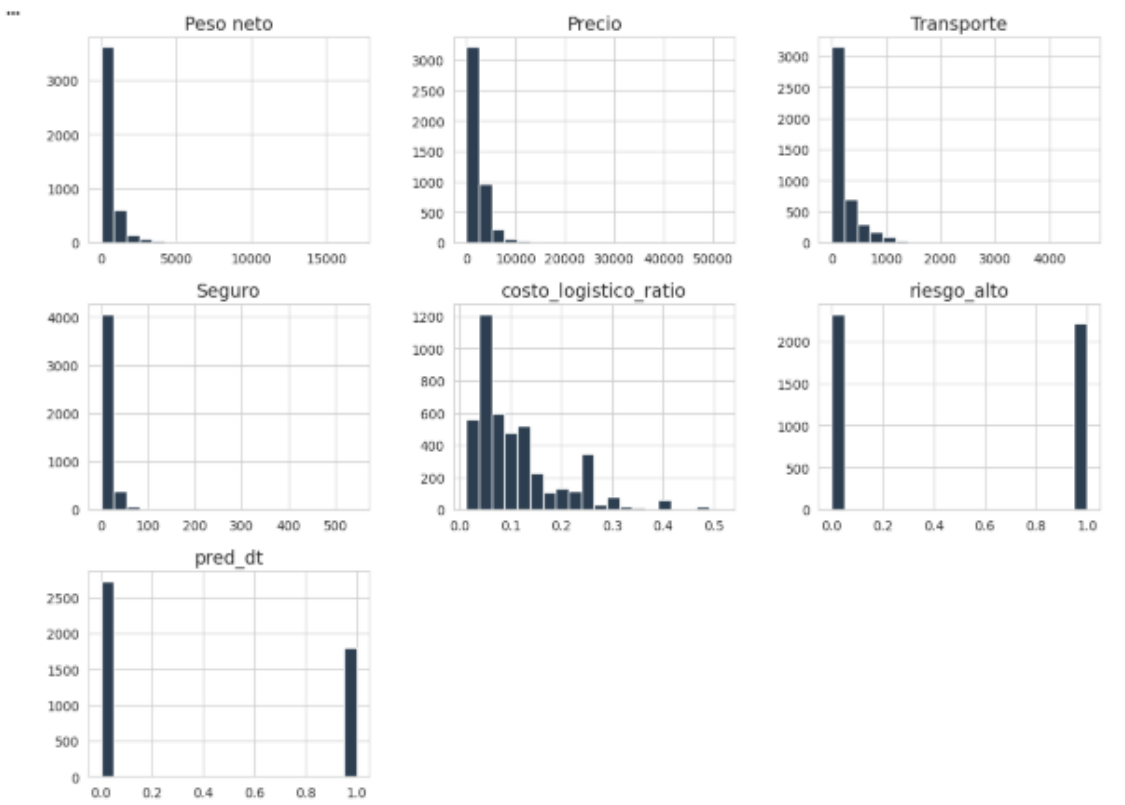
Distribución de las variables (histogramas)

Para suplementar los estadísticos descriptivos, se esbozaron histogramas de las primordiales variables numéricas, buscando discernir la

morfología distributiva, concentración de valores y evidencia de colas prolongadas o valores extremos. Esta visualización posibilita constatar asimetrías que, no invariablemente, se perciben solamente con promedios y medianas, a la vez que sustenta decisiones subsecuentes sobre percentiles, normas clasificatorias, y el trato a registros atípicos (Ver Tabla 6).

Figura 6

Histograma de variables numéricas



Los histogramas del Peso Neto, Precio, Transporte, y Seguro, exhiben una acentuada concentración en las escalas bajas y una cola alargada hacia valores superiores, un comportamiento coherente con transacciones disparejas, donde unas cuantas operaciones acumulan magnitudes, pesos, o costos logísticos significativos. El costo_logistico_ratio muestra la mayor frecuencia entre valores que rondan 0,03 a 0,15, con instancias menos frecuentes aproximándose a 0,5, esto delata operaciones en las que el transporte y el seguro conforman una porción elevada del precio total. Las

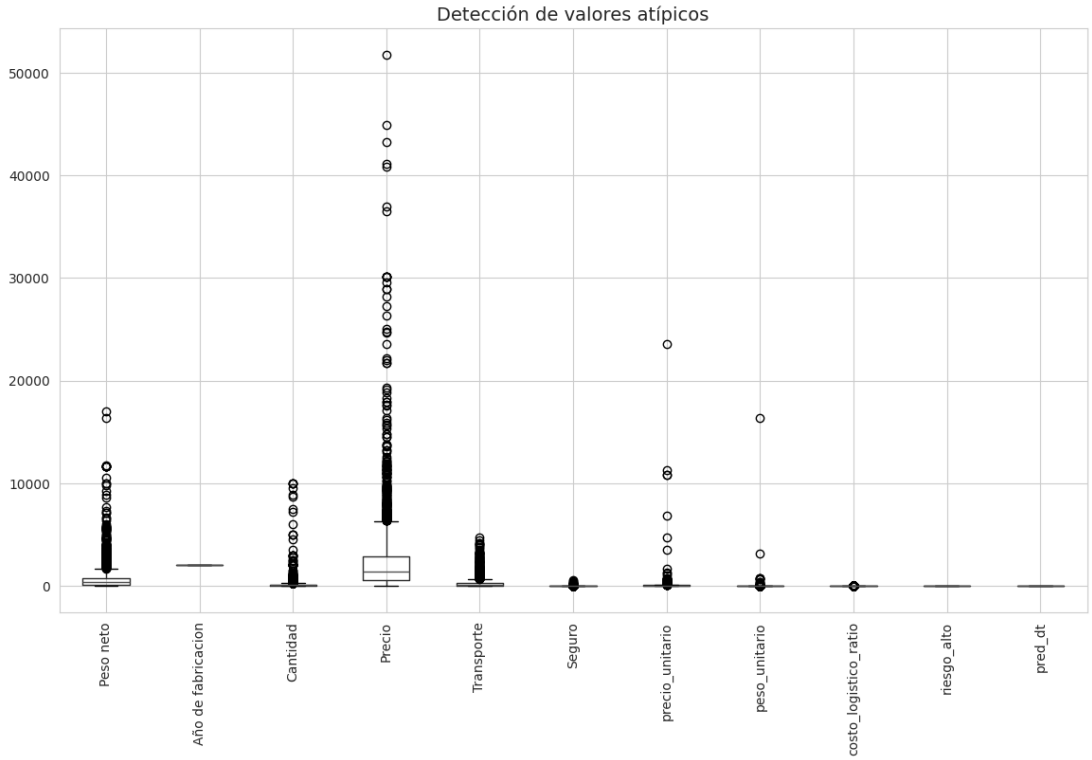
variables binarias riesgo_alto y pred_dt presentan dos barras preponderantes, una en 0 y otra en 1; riesgo_alto manifiesta una distribución casi equilibrada entre clases, mientras que pred_dt presenta más registros en 0, concordando con un árbol que cataloga menos transacciones como de alto riesgo en comparación con la etiqueta. Tales patrones justifican la adopción de medidas robustas tales como cuartiles e IQR, confirmando el subsiguiente paso analítico, donde se compararán las variables por clase, y se entrenará el modelo con métricas derivadas por unidad.

Boxplots (Detección de outliers)

Como complemento de los histogramas, se generaron diagramas de caja (boxplots) para identificar valores atípicos y comparar la dispersión de cada variable numérica en una misma gráfica. El boxplot de la Figura 7 resume mediana, cuartiles y rango intercuartílico y señala valores atípicos (fuera del rango habitual), información para establecer percentiles de control y evitar que extremos distorsionen medidas y modelos.

Figura 7

Boxplot para detección de outliers



En el gráfico se observan valores extremos en variables con escalas grandes. Precio agrupa la mayor cantidad de valores atípicos y el máximo superior a 50.000, que se corresponde con las mayores transacciones en un conjunto donde la mayoría se sitúa en precios bajos. Peso neto y Cantidad también son muy dispersos y con muchos valores atípicos lejos del cuerpo principal, lo que indica manipulaciones en lote o por unidades.

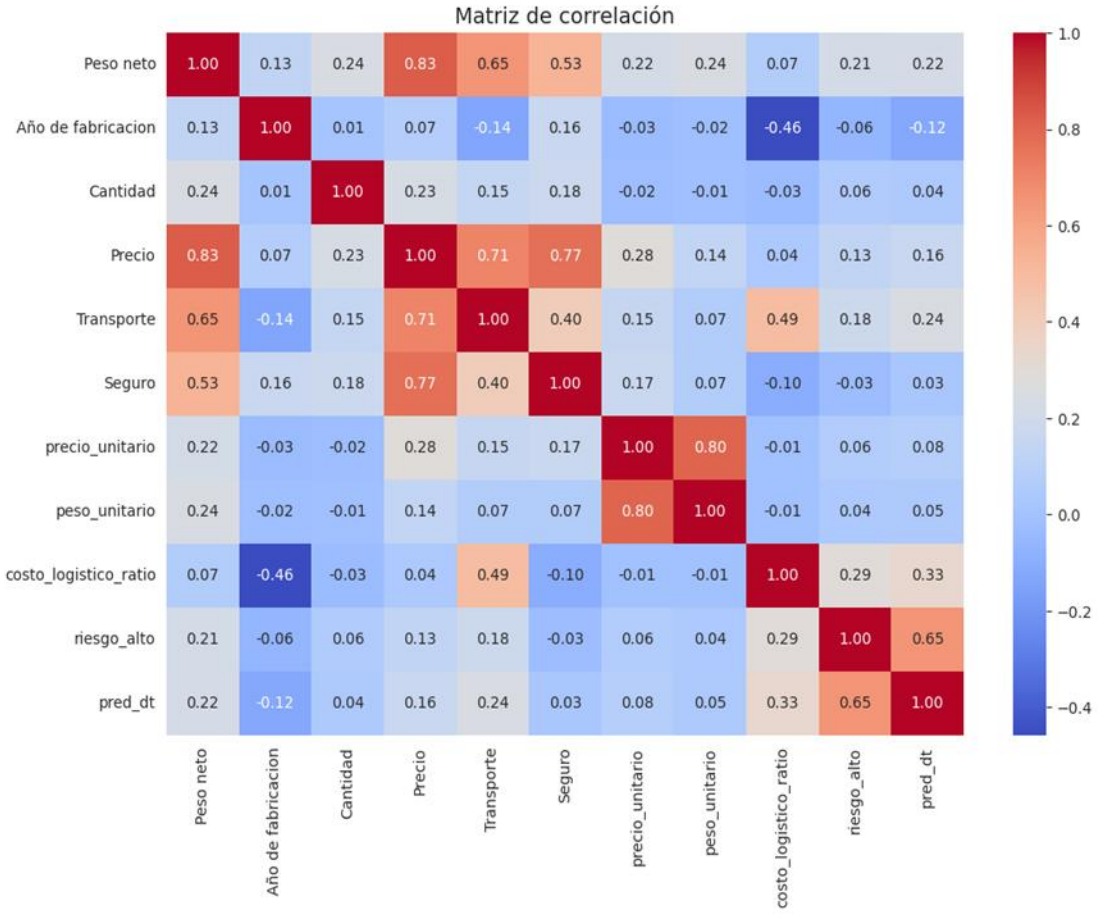
En Costos Asociados, Transporte y Seguro son valores extremos que pueden corresponder a fletes o pólizas fuera de lo común en relación al resto del grupo. En las variables calculadas, precio_unitario y peso_unitario tienen valores atípicos altos, consistentes con situaciones en las que alguna combinación de precio, cantidad o peso resulta en valores por unidad fuera de rango. costo_logistico_ratio: conserva una caja más plana, con algunos picos que corresponden a operaciones donde flete y seguro se llevan una mayor porción del precio.

Las variables binarias riesgo_alto y pred_dt están comprimidas en 0-1 (sin lectura de valores atípicos por ser variables dicotómicas). Esta revisión respalda el uso de cuartiles, IQR y percentiles como criterios numéricos antes del modelado.

Análisis de correlación

Para explorar relaciones lineales entre variables numéricas, se computó una matriz de correlación, y se visualizó en un mapa de calor. Esta técnica permite señalar variables que se relacionan entre sí, pares con vinculo inverso y probables redundancias entre predictores, información útil para descifrar el comportamiento del conjunto de datos y predecir variables con gran dependencia previo al entrenamiento.

Figura 8
Boxplot para detección de outliers



La matriz revela altas correlaciones entre variables de escala y costo: Peso neto-Precio (0,83): las más pesadas suelen ser las más caras; Precio-Seguro (0,77) y Precio-Transporte (0,71): el coste logístico crece con el precio. Transporte–Peso neto (0,65) y Seguro–Peso neto (0,53) refuerzan el mismo patrón físico–logístico. En variables calculadas, precio_unitario y peso_unitario están altamente correlacionadas (0,80), lo que tiene sentido en las transacciones dondelas unidades más pesadas tienden a tener mayor precio unitario.

El costo_logistico_ratio se correlaciona moderadamente con Transporte (0,49) y negativamente con Año de fabricación (-0,46), lo que sugiere que la razón logística aumenta cuando el transporte es mayor y tiende

a ser menoraumenta encuando el transporte es mayor y tiende a ser menor en los años más recientes del conjunto. En cuanto a la etiqueta, riesgo_alto se relaciona más con costo_logistico_ratio (0,29) que con el resto de variables, en línea con una regla de clasificación en la que la fracción logística interviene directamente. La predicción del árbol pred_dt se correlaciona con riesgo_alto (0.65), lo cual es esperable por ser una variable construida para aproximar la etiqueta, y por lo tanto no debe ser usada como variable explicativa.

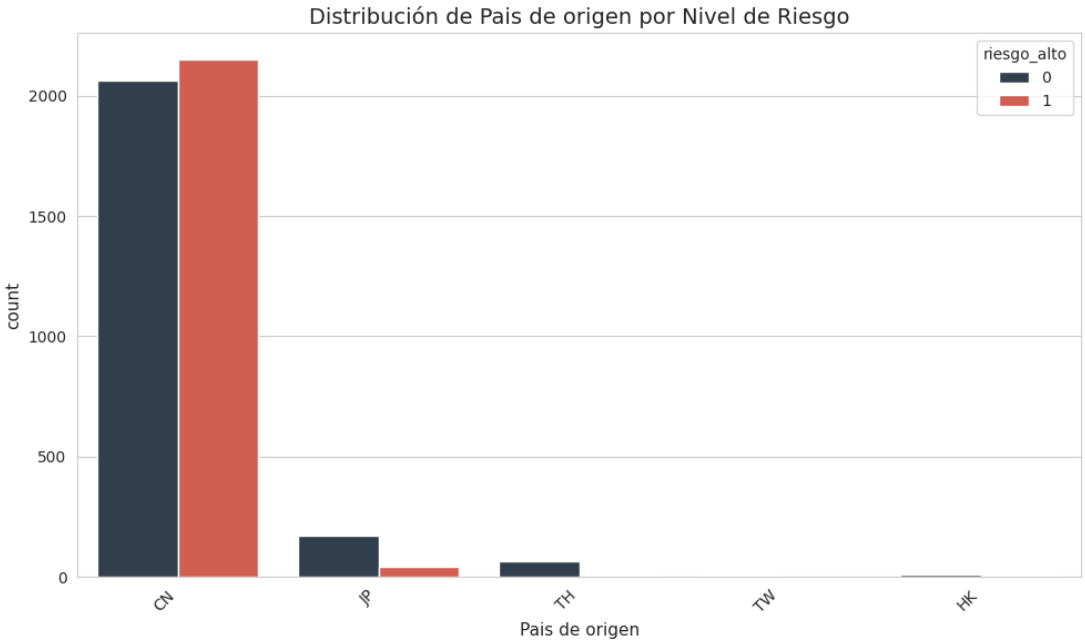
Concentración de categorías

Distribución por país

En esta parte se estudia la distribución de variables categóricas para identificar concentración por país de procedencia, tipo de mercancía y proveedores. Esta mirada permite hacerse una idea de qué tan balanceado está el conjunto de datos o si existe algún tipo de sesgo por sobreabundancia de ciertas categorías. También explica por qué las variables categóricas no añaden mucho al modelo cuando una categoría cubre la mayoría de los casos.

En la Figura 9 se muestra el país de origen de cada una de las transacciones realizadas.

Figura 9
País de origen



El gráfico ilustra la dispersión de operaciones por nación de origen, segmentada según la etiqueta riesgo_alto (0 = bajo, 1 = alto). La mayor parte de transacciones se origina desde China (CN), con elevadas cantidades en ambas categorías, evidenciando que China aglomera prácticamente todo el volumen de importaciones del conjunto de datos y, en consecuencia, concentra asimismo la mayor cantidad de situaciones categorizadas como riesgo alto.

En Japón (JP) y Tailandia (TH), la cifra de operaciones es significativamente más escasa; en estas naciones prevalece el bajo riesgo, y el alto riesgo se manifiesta con escasa periodicidad, o casi nula, eso mengua su impacto en el aprendizaje del modelo. Taiwan (TW) y Hong Kong (HK) exhiben una presencia insignificante, por lo que su aporte estadístico es exiguo y sus cotejos entre categorías no son consistentes debido a su reducida dimensión.

En cuanto a la interpretación, este diagrama manifiesta que el país de origen sirve más como una variable de segmentación que como un separador

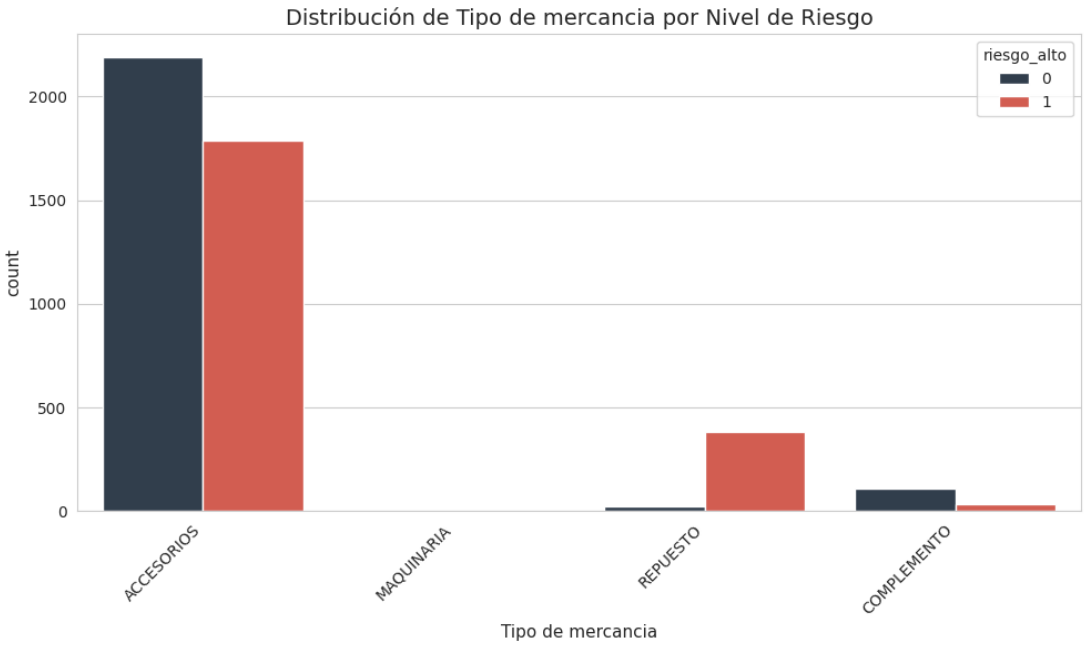
fundamental del riesgo, dado que la etiqueta se difunde dentro del país preponderante y las categorías minoritarias muestran pocas evidencias para discernir patrones con exactitud.

Distribución por tipo de mercancía

Para analizar las diferencias en el riesgo operativo según la naturaleza del producto importado, se compararon la cantidad de transacciones por tipo de mercancía, dividiendo las clases de riesgo alto (0 = bajo, 1 = alto). Visualizar esto permitirá ver, que categorías aglomeran el volumen de operaciones, además si alguna presenta mayor presencia relativa de alto riesgo, lo cual sería importante (Ver Figura 10).

Figura 10

Distribución por tipo de mercancía



ACCESORIOS, como ilustra el diagrama, domina las transacciones en ambos grupos, pero con más operaciones de bajo riesgo en cantidad. REPUESTO, por otro lado, evidencia una tendencia diferente, mostrando más casos de alto riesgo en comparación con los de bajo riesgo, lo que revela que esta clase está más propensa a las alertas.

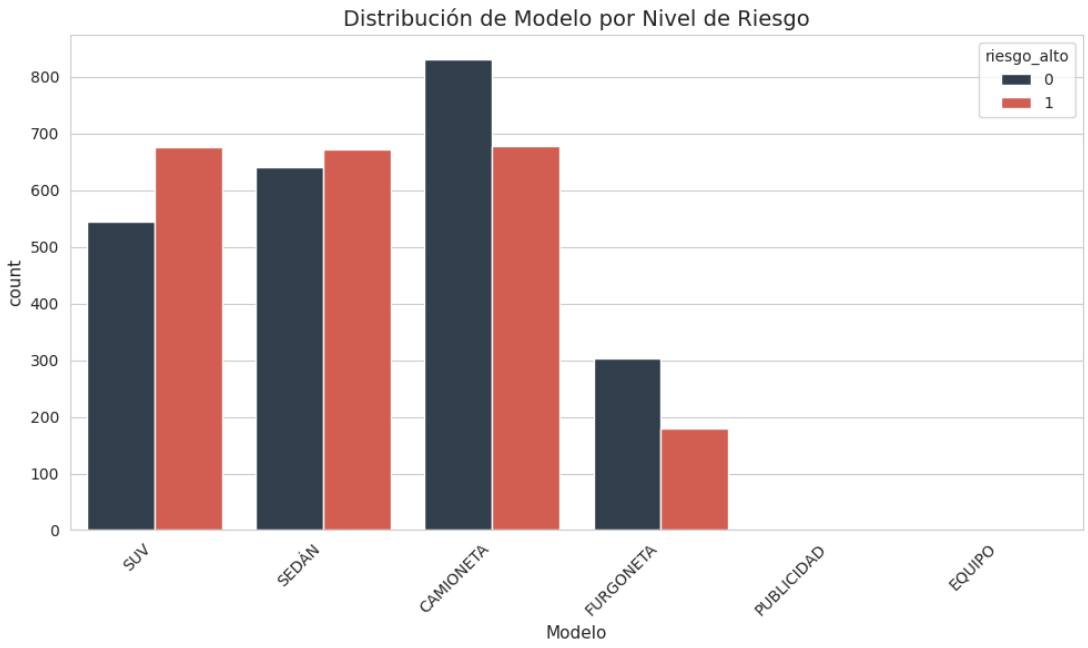
COMPLEMENTO presenta mayoritariamente bajo riesgo, mientras que el alto riesgo es menos frecuente, señalando un comportamiento más regular que REPUESTOS.

MAQUINARIA, casi invisible, impide cualquier comparación válida. Esta repartición valida el empleo del tipo de producto como factor de segmentación, enfocándose, sobre todo, en REPUESTO para optimizar las revisiones y controles antes de la compra.

Distribución del modelo por nivel de riesgo

Evaluar el riesgo operativo, se modifica por el modelo de transacción, pues se compararon las operaciones contadas por modelo, distinguiendo riesgo_alto (0 bajo, 1 alto). Así, se puede ver los modelos con mayor volumen y, si alguna categoría destaca, por su alto riesgo (Ver Figura 11).

Figura 11
Distribución del modelo por nivel de riesgo



El gráfico revela, observa, que CAMIONETA acapara la mayor parte de las transacciones. El riesgo bajo predomina claramente, lo que sugiere un considerable volumen de operaciones, dentro de los límites considerados buenos por la marca. En SUV y SEDÁN, sin embargo, los conteos de alto

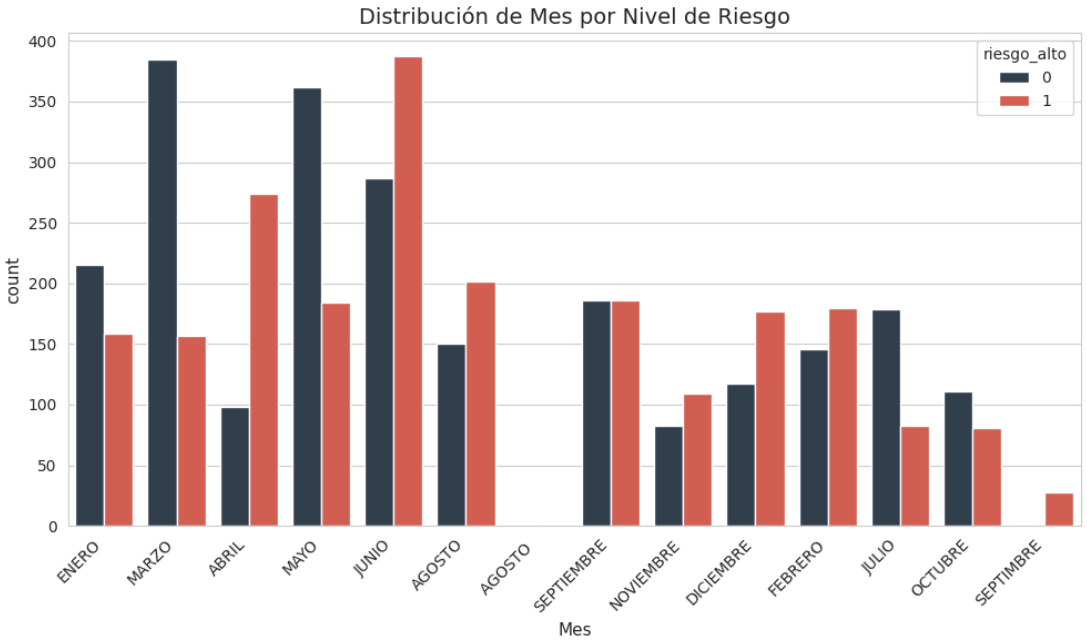
riesgo son parecidos o más altos que el bajo riesgo; esto muestra una distribución más centrada en las transacciones de alerta, en esas categorías.

FURGONETA muestra menor volumen, y un predominio del bajo riesgo. Con alto riesgo, pero, en menor proporción. Las categorías PUBLICIDAD y EQUIPO, parece que, casi no tienen registros, por lo que no permiten una comparación sólida. En cuanto a lo operativo, el modelo funciona como variable de segmentación. Las mayores alertas se agrupan en SUV y sedán, dentro del análisis. Mientras que camioneta aporta volumen con menos alto riesgo.

Distribución de mes por nivel de riesgo

Se comparó el número de transacciones mensuales, distinguiendo el riesgo operativo alto para examinar su evolución temporal. Analizando: riesgo_alto, donde 0 representa bajo y 1 alto. Esta gráfica ayuda a señalar los meses con mayor actividad general, como también aquellos meses con más operaciones de alto riesgo, esta información es valiosa para planificar los controles internos y priorizar las revisiones en momentos de mayor trabajo operativo (Ver Figura 12).

Figura 12
Distribución de mes por nivel de riesgo



En el gráfico se observa la variación mensual en la distribución de clases. Marzo y mayo son meses altos en operaciones y en su mayoría y en su mayoría de bajo riesgo, lo que significa que son meses con muchas operaciones en rangos aceptables. Junio se caracteriza por un aumento del alto riesgo sobre el bajo riesgo, y abril también tiene una alta de concentración operaciones de alto riesgo en relación con las de bajo riesgo. En septiembre se igualan las clases, pero noviembre, diciembre y febrero siguen siendo de alto riesgo. Julio: más bajo riesgo y menos alto riesgo. El comportamiento evidencia que el riesgo no es homogéneo a lo largo del año y que hay meses que acumulan más alertas, lo que permite programar revisiones documentales, validación de costos logísticos y controles por proveedor en los meses de mayor riesgo.

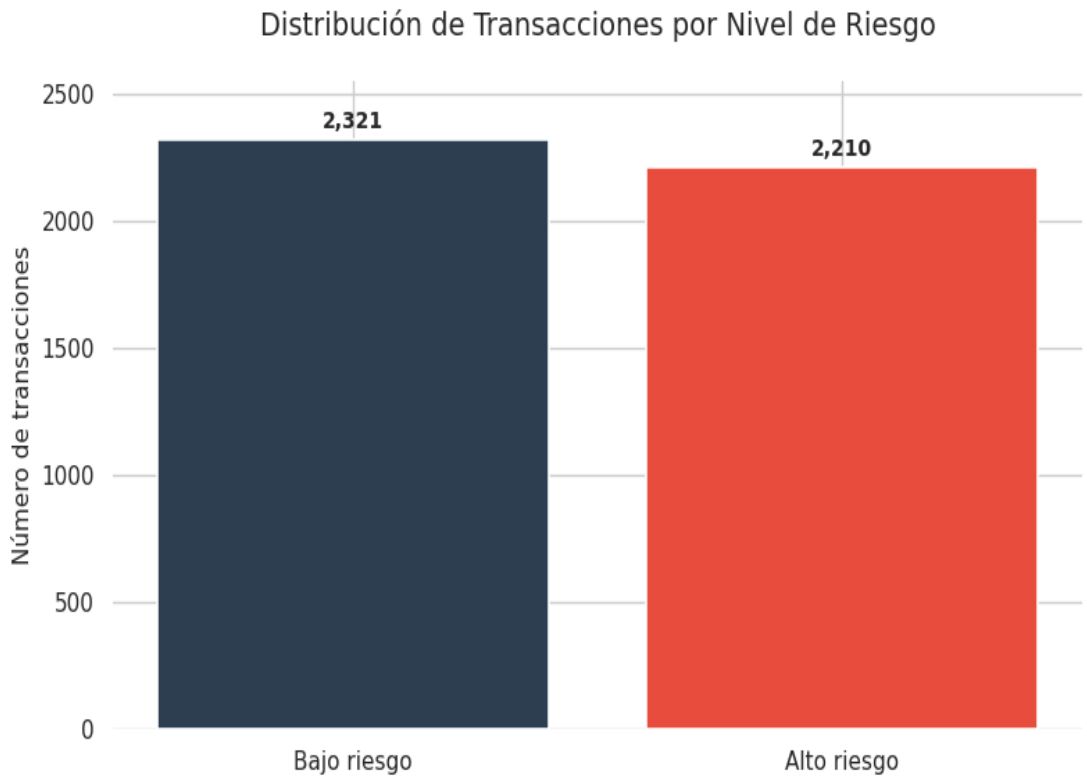
Ejecución del modelo predictivo y evaluación de resultados

Los resultados del modelo predictivo que se orienta a prevenir riesgos económicos en PYMES importadoras se muestran en esta sección. Dado que este paso posibilita analizar la proporción de casos clasificados como "alto" y

"bajo" riesgo en el total de transacciones, la revisión básica de la distribución de clases producidas por la etiqueta de riesgo es el primer paso del análisis. La lectura ayuda a entender el nivel de exposición operativa y la carga de trabajo preventiva que el modelo está detectando en la base analizada. En la Figura 13 se plasma la distribución de transacciones por nivel de riesgo

Figura 13

Distribución de Transacciones por Nivel de Riesgo



Se muestra en la Figura 1 la repartición de las transacciones según su nivel de riesgo. Se contabilizan 2.321 transacciones de bajo riesgo y 2.210 de alto riesgo, lo que equivale a un total de 4.531. El 48,78% representa el riesgo alto, mientras que el bajo riesgo supone el 51,22%. Esta proporción de aproximadamente 50/50 indica que las variables derivadas y la regla de etiquetado están dividiendo el conjunto en dos grupos de tamaño similar, lo cual minimiza el sesgo hacia una única clase y permite un entrenamiento más sencillo para los clasificadores. Un 48,78% de alto riesgo desde la lectura operativa indica que cerca de la mitad de las transacciones necesitarían ser revisadas conforme a normas de control, priorización de verificaciones y

supervisión de costos logísticos, precios unitarios extremos o concentración por proveedor.

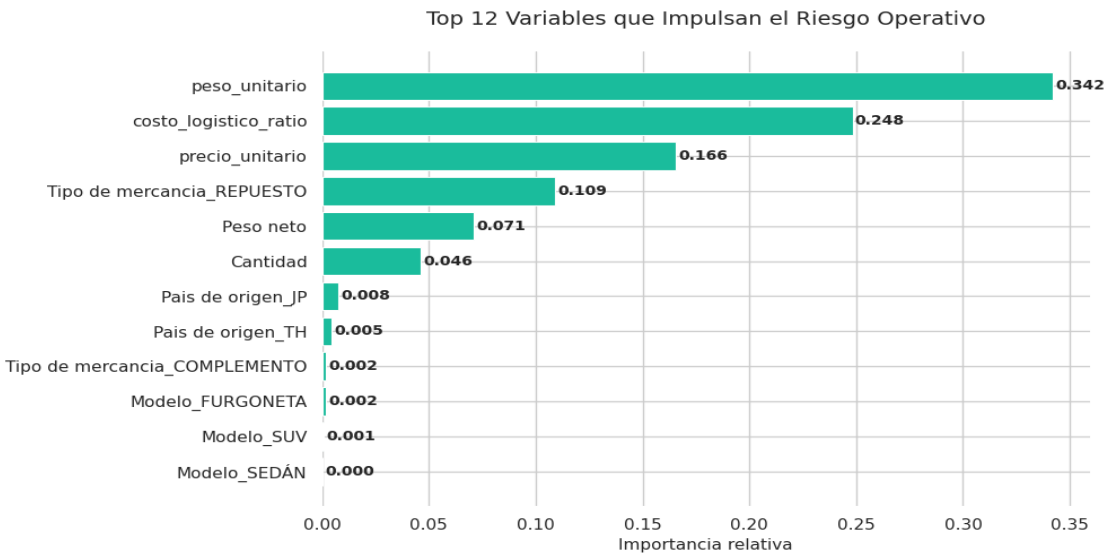
El paso siguiente, después de esta distribución general, es determinar qué variables están impulsando la clasificación y en qué medida, para convertir el resultado del modelo en criterios prácticos de control financiero. Por tanto, el estudio prosigue con el gráfico de la importancia de las variables.

Variables impulsoras del riesgo operativo

Se examina la relevancia de las variables que Random Forest ha calculado para poder profundizar en el resultado del modelo y convertirlo en parámetros de control financiero. Esta salida clasifica los predictores de acuerdo a su contribución relativa en el establecimiento de reglas internas del conjunto de árboles. La lectura se lleva a cabo como un ranking de contribuciones dentro del modelo entrenado, lo cual es útil para determinar qué indicadores monitorear más frecuentemente en el proceso de importación. En la Figura 14 se clasifican los predictores que más aportan al modelo y se estudian cuáles explican de manera más efectiva la asignación del riesgo.

Figura 14

Top de variables que impulsan el riesgo operativo



La imagen muestra que el peso_unitario (0,342), el costo_logístico_ratio (0,248) y el precio_unitario (0,166) representan 0,756 del peso total, lo que corresponde al 75,6% de la relevancia total.

Desde un Punto de vista técnico, el modelo se basa principalmente en variables por unidad y en una relación logística: si aumenta el costo logístico relativo (seguro + transporte / precio), si el peso por unidad crece o si hay un desplazamiento del precio unitario hacia valores extremos, el algoritmo tiene más capacidad para distinguir entre transacciones de alto riesgo y bajo riesgo.

Aparecen en un segundo nivel el tipo de mercancía, el peso neto y la cantidad, con valores de 0,109, 0,071 y 0,046 respectivamente. Juntos constituyen un 22,6% adicional; lo cual indica que tanto el tipo de mercancía como el tamaño de la operación son útiles para mejorar la clasificación.

Los valores de las variables categóricas codificadas por país y modelo son bajos (cada uno ≤0,008), lo que sugiere que, en esta base, el patrón de riesgo se fundamenta más en magnitudes económicas y logísticas que en las etiquetas de origen o tipo de vehículo.

Una vez establecida esta jerarquía, el siguiente paso es comprobar cuán efectivamente ambos modelos separan las clases en términos de errores y aciertos. Por lo tanto, el análisis sigue con la matriz de confusión comparativa (Random Forest vs Árbol de Decisión), que examina los falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos para guiar las acciones de prevención y priorización de revisión.

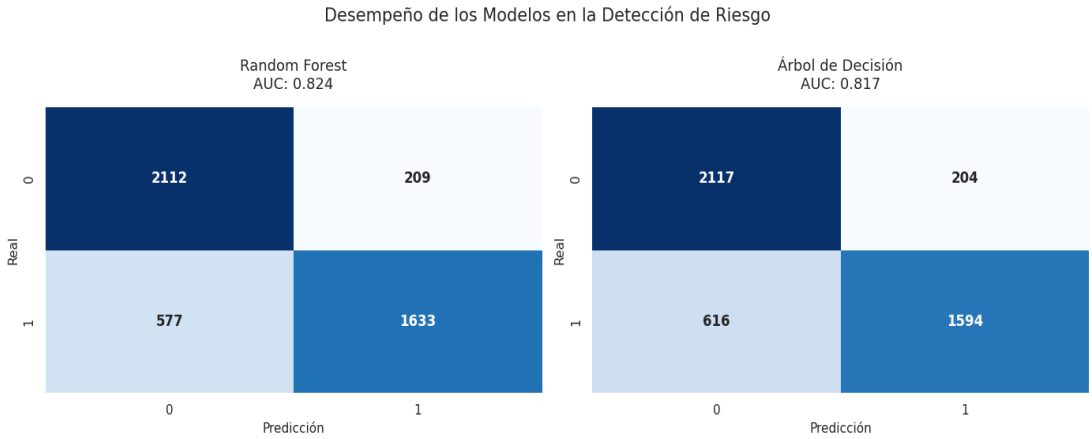
Desempeño de Modelos de detección de riesgos mediante matriz de confusión combinada

Para determinar si el modelo contribuye a evitar riesgos financieros, se examina su competencia para clasificar adecuadamente las transacciones de bajo riesgo (clase 0) y aquellas de alto riesgo (clase 1). La Figura 15 muestra para cada uno de los algoritmos, una matriz de confusión que incluye cuatro conteos: verdaderos positivos (1➔1), falsos negativos (1➔0), falsos positivos

(0➔1) y verdaderos negativos (0➔0). En las tareas de alerta temprana, es importante monitorear con particular atención los falsos negativos, ya que representan casos peligrosos que el modelo clasifica como de bajo riesgo.

Figura 15

Top de variables que impulsan el riesgo operativo



Se puede ver en Random Forest un total de 2.112 verdaderos negativos y 1.633 verdaderos positivos, además de 577 falsos negativos y 209 falsos positivos. Con estos valores, el rendimiento global llega a un 82,65% de exactitud y a un 88,65% de precisión (cuando el modelo indica "alto riesgo", suele atinar) y a un 73,89% de recuperación (detecta aproximadamente tres cuartas partes de los casos reales que tienen alto riesgo). La especificidad es del 90,99%, lo que indica un control adecuado de falsos positivos en la categoría de bajo riesgo; el F1 es 0,806, una relación equilibrada entre precisión y recuperación. El diagrama muestra que el AUC es igual a 0,824; este número refleja la capacidad de separación entre las clases y respaldado la lectura comparativa del rendimiento.

En Árbol de Decisión se obtuvo 2.117 verdaderos negativos y 1.594 verdaderos positivos, 204 falsos positivos y 616 falsos negativos. La exactitud es de 81,90%, la precisión se mantiene en 88,65%, el recuerdo disminuye a 72,13%, la especificidad es de 91,21% y el F1 es de 0,795. En comparación, Random Forest disminuye los falsos negativos en 39 casos (577 vs 616) y aumenta los verdaderos positivos en la misma cantidad (1.633 vs 1.594), lo

cual es preferible si se busca identificar más transacciones de alto riesgo para revisión. El AUC también es marginalmente superior en Random Forest (0,824 vs 0,817).

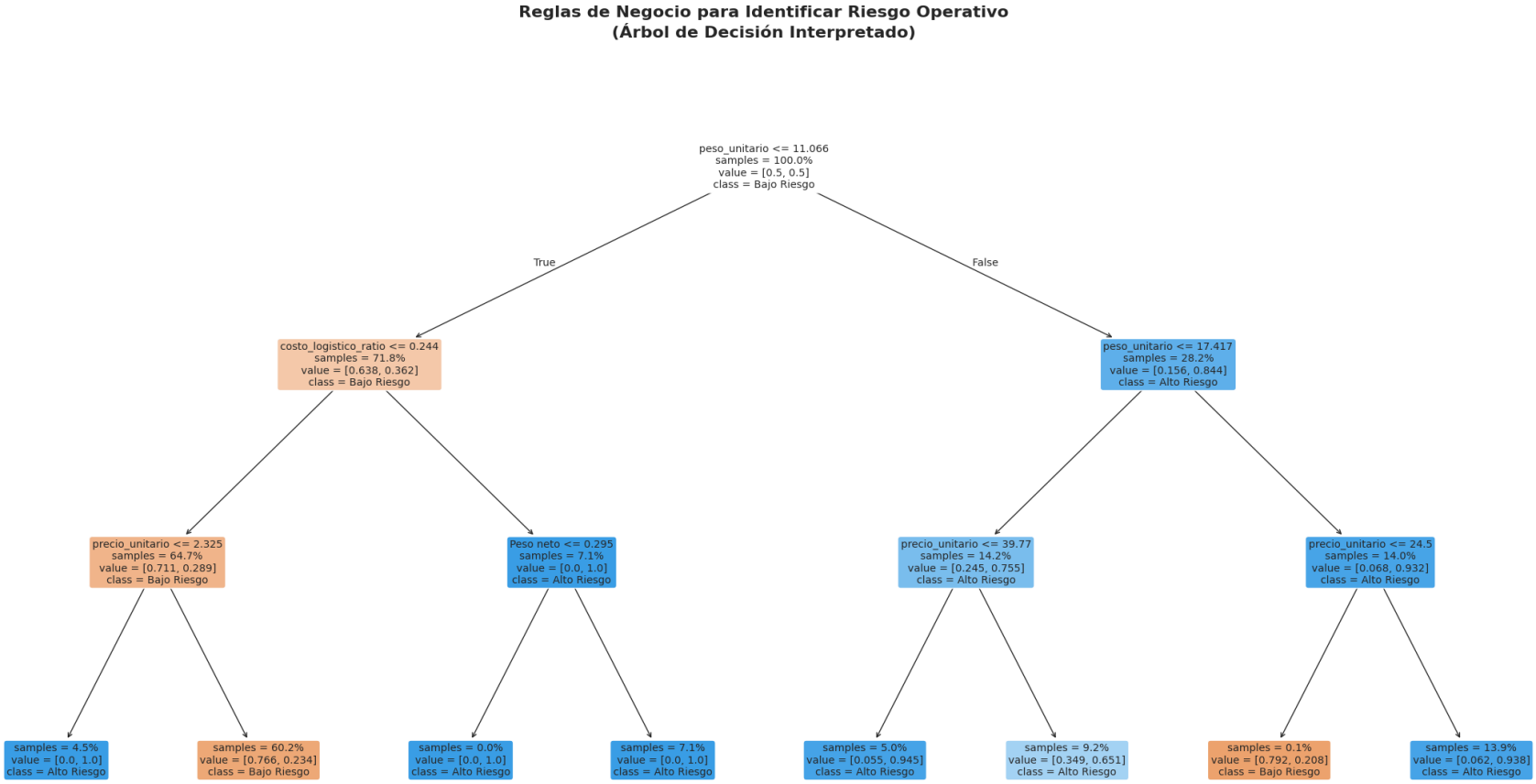
Con esta aprobación, el proceso sigue con la interpretación del Árbol de Decisión, donde se pueden observar las reglas tipo "si-entonces" que definen por qué una transacción se considera de alto o bajo riesgo, lo que sirve para documentar criterios operativos y estandarizar revisiones internas.

Reglas de negocio para identificar riesgos operativos a través de Árbol de Decisión

Para transformar el resultado del modelo en normas que se puedan aplicar al control interno, se examina el Árbol de Decisión como un grupo de reglas del tipo "si-entonces". Cada nodo de este diagrama representa la variable utilizada para dividir (por ejemplo, *precio_unitario*, *costo_logistico_ratio* o *peso_unitario*), el umbral numérico, el porcentaje aproximado de muestras que alcanzan ese punto (*samples*), la proporción por clase (*value*) y la clase asignada a ese nodo. Las ramas simbolizan elecciones binarias: la condición (≤) se cumple a la izquierda, y no se cumplea la derecha (>). Esta lectura posibilita la justificación de por qué una transacción se convierte en "alto" o "bajo" riesgo usando criterios que pueden ser medidos (Ver Figura 16)

Figura 16

Reglas de Negocio para Identificar Riesgos Operativos



La primera regla surge de que el *peso unitario* es menor o igual a 11.066, lo que hace que el árbol clasifique como riesgo bajo. En esa área, el árbol verifica que el *costo_logístico_ratio* no supere 0.244; si el ratio logístico es bajo y el *precio_unitario* está por debajo de 2.325, la mayoría de las situaciones permanecen con un riesgo bajo. Por otro lado, si el índice logístico es mayor a 0.244 o si se presentan combinaciones en las que el árbol conduce a nodos con la clase "alto riesgo", la transacción queda marcada para ser revisada.

Cuando el *peso_unitario* es superior a 11.066, el árbol se vuelve más conservador y clasifica en su mayoría como alto riesgo, afinando la decisión con recortes extra de *precio_unitario* (como ≤ *39.77* o ≤ *24.5*).

En términos operativos, el árbol indica que el riesgo se incrementa si el peso por unidad se eleva o si el precio unitario se encuentra en los rangos asociados por el modelo con una probabilidad de alerta más alta, utilizando como condición adicional la razón logística.

El análisis subsiguiente, con estas reglas ya especificadas, transfiere el resultado del modelo a una salida de gestión: la clasificación de riesgo por proveedor. En ella se agrupan las transacciones por proveedor, se calcula la tasa de riesgo y se otorgan prioridades a los proveedores para realizar auditoría, renegociar condiciones o revisar documentos.

Ranking de riesgo por proveedor

Para convertir el resultado del modelo en un análisis operativo, se crea una clasificación de proveedores según la cantidad de transacciones documentadas. La Tabla 2 posibilita identificar a aquellos proveedores que concentran un volumen más alto de operaciones en la base analizada, lo que simplifica establecer prioridades para el control y la revisión, puesto que una cantidad mayor de transacciones eleva la exposición acumulativa a eventos riesgosos dentro del flujo importador.

Tabla 3
Ranking de volumen transaccional por proveedor

🏆	Proveedor	total_transacciones
1	ANHUI AGGEUS AUTO -TECH CO., LTD	3
8	FOSHAN YONGMINGSHANG TRADING COMPANY	5
4	DANYANG DEBAO AUTO SPARE PARTS CO., LTD	4
30	SHENZHEN ZEROELEC TECHNOLOGY CO., LTD	2
29	SHENZHEN RUNBO LED CO., LIMITED	4
40	ZHEJIANG GUAN AN INTELLIGENT FIRE FIDHTING TEC.	1
36	WENZHOU NOBLE AUTO PARTS CO., LTD	1
26	SHANGHAI ANZOO AUTOPARTS INDUSTRY CO., LTD	31
19	NANJING WINLAND HOME BUILDING CO., LTD	21
22	POTENTIAL SMARTTOP CO., LTD	6

La Tabla 2 presenta diversidad en la participación de cada proveedor. Nanjing Winland Home Building Co., Ltd. y Shanghai Anzoo Autoparts Industry Co., Ltd. son ejemplos de proveedores que tienen una concentración elevada de registros, con 21 y 31 transacciones respectivamente, cifras que sobrepasan significativamente las del resto del grupo mostrado.

FOSHAN YONGMINGSHANG TRADING COMPANY (5) y POTENTIAL SMARTTOP CO., LTD (6) representan la segunda categoría de actividad, y otros proveedores muestran niveles bajos (de 1 a 4). Este patrón indica que el análisis de riesgo por proveedor, realizado posteriormente, no solo debe tener en cuenta la proporción o tasa de riesgo, sino también el volumen de transacciones y el monto relacionado. Esto se debe a que un proveedor con un gran número de operaciones puede generar una carga preventiva mayor incluso con tasas moderadas.

Una vez que se ha determinado el ranking por volumen, el siguiente paso consiste en combinar estos datos con la clasificación del modelo para calcular la tasa de riesgo por proveedor y elaborar una priorización más exacta. El próximo resultado examina a los proveedores que presentan la

mayor cantidad de transacciones etiquetadas como de alto riesgo, así como el total de transacciones y el gasto total, con el fin de guiar las acciones de monitoreo y control.

Según dos indicadores, la tasa de riesgo y el gasto total, la Tabla 3 muestra el orden jerárquico de los proveedores. La tasa resume el porcentaje de transacciones que se consideran de alto riesgo para cada proveedor, mientras que la cifra total de gastos es un cálculo aproximado del volumen económico correspondiente. Esta lectura posibilita la combinación de magnitud monetaria, proporción y frecuencia para establecer prioridades de control.

Tabla 4

Ranking de volumen transaccional por proveedor

🏆	Tasa_riesgo	Gasto_total
1	1.0	1170.00
8	1.0	14070.00
4	1.0	122631.00
30	1.0	45780.00
29	1.0	86294.12
40	1.0	22000.00
36	1.0	21760.00
26	1.0	116868.00
19	1.0	16956.31
22	1.0	32500.00

En los diez ejemplos presentados, la *tasa_riesgo* es igual a 1.0; esto indica que se clasificaron como de alto riesgo todas las operaciones de cada proveedor pertenecientes a este subconjunto. Conforme a este resultado, la distinción operativa se traslada al gasto total, debido a que el volumen monetario presenta una amplia variación: sobresalen SHANGHAI ANZOO AUTOPARTS INDUSTRY CO., LTD (116.868,00), DANYANG DEBAO AUTO SPARE PARTS CO., LTD (122.631,00) y SHENZHEN RUNBO LED CO., LIMITED (86.294,12) como los proveedores con el mayor monto expuesto dentro del grupo listado.

A pesar de que los montos son más bajos, otros proveedores conservan riesgo total en su portafolio de transacciones, como ANHUI

AGGEUS AUTO -TECH CO., LTD (1.170,00). En lo que respeta a la priorización, una regla práctica consiste en comenzar con los proveedores cuyo gasto es más elevado y cuya tasa es 1.0, porque estos presentan una alta concentración de alertas junto con un mayor valor económico comprometido.

A continuación, se muestra el conjunto de datos finales después de la depuración, transformación y creación de variables, para confirmar que los datos empleados por el modelo sean coherentes y posibiliten la interpretación de resultados en términos de transacción.

La Tabla 4 combina campos descriptivos (como el país, la mercancía, el proveedor, el modelo y el periodo) con magnitudes físicas y montos. Esto posibilita que se pueda examinar la coherencia entre registros y la trazabilidad del proceso de construcción de los indicadores que sustentan la clasificación de riesgo.

Tabla 5

Ranking de riesgo por proveedor

Índice	Descripción	Pais de origen	Tipo de mercancía	Peso neto	Modelo	Mes	Año de fabricación	Cantidad	Precio	Transporte	Seguro	Proveedor	precio_unitario	peso_unitario	costo_logistico_ratio	riesgo_alto	pred_dt
0	Decorativo Flores Pvc	Cn	Accesorios	2364.23	Suv	Enero	2021	21.0	2100.0	718.55	32.385	Shenzhen Mind Trading Co.,Ltd	100.0	112.582381	0.357588	1	1
1	Bolsas Pvc	Cn	Accesorios	2289.18	Sedán	Enero	2021	61.0	1952.0	695.74	31.357	Shenzhen Mind Trading Co.,Ltd	32.0	37.527541	0.372488	1	1
2	Peliculas Pvc	Cn	Accesorios	1313.46	Camioneta	Enero	2021	70.0	1400.0	399.19	17.992	Shenzhen Mind Trading Co.,Ltd	20.0	18.763714	0.297987	1	0
3	Aditivo Pvc	Cn	Accesorios	422.18	Suv	Enero	2021	10.0	450.0	128.31	5.783	Shenzhen Mind Trading Co.,Ltd	45.0	42.218000	0.297984	1	1
4	Aditivo Pvc	Cn	Accesorios	2110.92	Suv	Enero	2021	50.0	1300.0	641.56	28.916	Shenzhen Mind Trading Co.,Ltd	26.0	42.218400	0.515751	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
4526	Faro De Carretera	Cn	Complemento	44.64	Suv	Diciembre	2025	12.0	747.6	21.62	7.690	Dlaa Industrial Co., Ltd	62.3	3.720000	0.039205	0	0

Índice	Descripción	Pais de origen	Tipo de mercancía	Peso neto	Modelo	Mes	Año de fabricación	Cantidad	Precio	Transporte	Seguro	Proveedor	precio_unitario	peso_unitario	costo_logistico_ratio	riesgo_alto	pred_dt
4527	Faro De Carretera	Cn	Complemento	69.16	Camioneta	Diciembre	2025	36.0	1440.0	41.64	14.820	Dlaa Industrial Co., Ltd	40.0	1.921111	0.039208	0	0
4528	Faro De Carretera	Cn	Complemento	133.20	Furgoneta	Diciembre	2025	72.0	2448.0	70.79	25.190	Dlaa Industrial Co., Ltd	34.0	1.850000	0.039208	0	0
4529	Faro De Carretera	Cn	Complemento	226.45	Sedán	Diciembre	2025	24.0	1200.0	34.70	12.350	Dlaa Industrial Co., Ltd	50.0	9.435417	0.039208	0	0
4530	Faro De Carretera	Cn	Complemento	111.00	Suv	Diciembre	2025	72.0	2448.0	70.79	25.190	Dlaa Industrial Co., Ltd	34.0	1.541667	0.039208	0	0

El conjunto tiene 17 variables y 4.531 transacciones, entre las cuales sobresalen tres métricas derivadas que normalizan la lectura del riesgo: costo_logistico_ratio, peso_unitario y precio_unitario. Se nota en las primeras anotaciones que el peso_unitario es muy alto (por encima de 100, por ejemplo) y también el costo_logístico_ratio es elevado (más de 0.25), lo cual hace que la etiqueta riesgo_alto=1 se active. Al final de la tabla, las transacciones de FARO DE CARRETERA muestran un precio unitario moderado y un peso unitario bajo, además de una relación costo_logístico cercana a 0,039. Esto coincide con que tanto el riesgo alto como la predicción DT son iguales a cero.

Esta comparación entre los extremos evidencia que el criterio de riesgo se comporta de manera coherente con la lógica económica del indicador: los valores unitarios inusuales y la elevada carga logística relativa suelen estar asociados con alerta, en tanto que las relaciones logísticas bajas tienden a reunirse en bajo riesgo.

Se nota además un aspecto técnico que necesita ser documentado: hay, como mínimo, un caso con pred_dt=0 y riesgo_alto=1 (como la fila 2 del extracto), lo cual significa un falso negativo para el árbol. Desde el punto de vista operativo, este tipo de error es el más delicado cuando la meta es prevenir, ya que una transacción que las reglas internas han marcado como riesgosa queda sin alerta.

La existencia de estos casos hace relevante que el análisis no se restrinja a la etiqueta final, sino que examina los errores por segmento y sus causas, en particular cuando se trata de bienes y proveedores de mayor volumen o con un gasto asociado más alto.

Una vez examinada la distribución de clases, la importancia de las variables y el potencial de clasificación de los modelos, se añade un indicador sintético que sintetiza el grado de exposición del grupo examinado.

KPI Riesgo de quiebra

Este KPI convierte la categorización de transacciones por transacción en una proporción general que sirve para informar sobre la situación del riesgo

y establecer criterios de priorización para el control financiero y el seguimiento operativo.

KPI de Riesgo Operativo: 48.78%

- Transacciones de alto riesgo: 2,210

- Total de transacciones: 4,531

- Nivel de severidad: Alto

El KPI del riesgo operacional es 48.78% y se calcula a partir de 2,210 transacciones de alto riesgo sobre un total de 4,531, con un nivel de severidad alto. Este valor significa que, por lo general, una de cada dos operaciones queda marcada para revisión conforme a las reglas y el clasificador usado.

Este KPI propone tres medidas directas para prevenir problemas financieros:

(1) dar prioridad a los controles de proveedores y mercancías con un gasto total más alto,
(2) implementar validaciones previas en transacciones que tengan una relación costo/ logístico alta y valores unitarios extremos,
(3) conservar el seguimiento del KPI mes a mes y cotejarlo por período para detectar aumentos específicos. Un KPI de esta envergadura también permite la modificación de las reglas de alerta y los umbrales, para así enfocar las revisiones en aquellos casos con una mayor exposición económica y disminuir la carga de revisión en operaciones que tienen un bajo impacto monetario.

Calculadora de probabilidad de riesgo

El propósito es mostrar dos ejemplos opuestos que el modelo predice como Alto Riesgo con alta probabilidad y Bajo Riesgo con baja probabilidad, usando exactamente las mismas variables con las que se entrena el Árbol de Decisión. En la Tabla 5 se muestran los insumos principales, métricas derivadas calculadas para lo que corresponde al precio unitario, el costo unitario y costo logístico para cada uno de los procesos logísticos.

Tabla 6
Ranking de riesgo por proveedor

Caso	Proveedor	Pieza (Descripción)	Precio	Transporte	Seguro	Cantidad	Peso Neto	Resultado (Calculadora)
1 (Bajo)	KARPRO AUTO INDUSTRY INC.	TAPA DE BALDE	10,00	0,25	0,10	1	26,00	**Bajo Riesgo** (Prob: **20,80%**)
2 (Alto)	YIWU CITY MINGLU EXPORT IMPORT CO.LTD	PERILLA PALANCA DE CAMBIOS	362,00	4,02	3,66	200	28,97	**Alto Riesgo** (Prob: **100,00%**)

En el Caso 1, la razón logística es baja

$$\frac{0{,}25 + 0{,}10}{10} = 0{,}035$$

Y el precio unitario se mantiene en un rango operativo esperado (10,00), por lo que el árbol asigna Bajo Riesgo con probabilidad baja.

En el Caso 2, el precio unitario cae en un valor extremo inferior

$$\frac{362}{200} = 1{,}81$$

Condición que activa una regla de alerta del árbol y produce Alto Riesgo con probabilidad alta. Esta prueba permite utilizar la calculadora como un filtro previo a la aprobación: se introduce una compra planteada, se clase obtiene y probabilidad, y se determina revisión documental, verificación de costos logísticos o validación de registro antes del pago o nacionalización.

DISCUSIÓN

El objetivo específico de interpretación de resultados se logra al transformar las salidas del modelo en criterios perceptibles para la gestión del riesgo financiero en importadoras PYMES. En este artículo se modela una clasificación binaria de “alto” y “bajo” riesgo a partir de datos históricos operativos y financieros de transacciones, creando variables por unidad y una razón logística que resume el peso de transporte y seguro sobre lo comprado. En el modelado, la etiqueta riesgo_alto es como una alerta precoz ante problemas económicos operativos. Implica gastos logísticos enormes en proporción a lo adquirido o precios unitarios incorrectos. Así, esa señal permite centrar la atención en la revisión y control, justo antes de que el peso de los costos afecte la liquidez o los pagos. Esta intuición se alinea con la literatura de predicción de quiebra e implicaciones en el rendimiento financiero desarrollado por Gajdosikova y Valaskova (2023), donde las primeras señales se encuentran en medidas derivadas, estabilidad de márgenes, presión de costos o sensibilidad a factores externos, en lugar de una cifra aislada.

Un ejemplo de ello son los estudios que comparan modelos y demuestran superioridad operativa de Random Forest en métricas generales de clasificación. Por ejemplo, un estudio realizado por Samara y Shinde (2025) que compara modelos de predicción de quiebra basados en datos contables encuentra que los modelos de ensamble generalmente obtienen resultados comparables en precisión y F1 a modelos más sencillos. En este estudio, el AUC se mantiene alrededor de 0.82, lo que justifica el uso de conjuntos como Random Forest para modelar relaciones no lineales entre variables operativas. A su vez, el Árbol de Decisión sigue siendo valioso por su habilidad para representar la clasificación en reglas de control, las cuales pueden ser implementadas en procesos internos.

Otro punto interesante es cómo interpretar el riesgo en shocks como los de pandemia y post-pandemia. Un estudio de Bozkurt y Veysel (2023) usa Random Forest con explicabilidad tipo SHAP para descubrir factores relacionados con mayor probabilidad de angustia y quiebra, señalando que acceso a financiamiento, antigüedad y endeudamiento aumentan el riesgo.

Aunque este análisis no incluye deuda ni financiación directa, refleja un aspecto funcional de interés impactando las ganancias el precio logístico proporcional y la estructura por unidad. Cuando logística y cobertura significan una parte notable del precio pagado, la ganancia operacional se ve comprimida, obligando a la empresa a mayor liquidez, esto para mantener el inventario y aduanas, incrementando el riesgo a estrés económico.

Un primer resultado operativo es el tamaño del riesgo que se ve en la base. El estudio encuentra que 48.78% de las transacciones se clasifican como “alto riesgo” según el criterio construido, es decir, casi la mitad de la actividad, lo que exige pensar en priorización, muestreo de auditoría y reglas de control que no saturen la operación. En riesgo, un índice alto no significa “quiebra inminente”, sino potencial sobrecarga de revisión: justificación de costos logísticos, precios unitarios extremos, consistencia de cantidades, dependencia por proveedor. Pero en la práctica, un gran número de alertas siempre requiere establecer prioridades. La revisión se puede disparar por umbrales: ratio logístico por encima del valor de control, precio unitario fuera de percentiles o proveedores con altas tasas de riesgo. También se puede usar muestras estratificadas por importación, para que las transacciones de menor importación no gasten los mismos esfuerzos que las de mayor riesgo monetario. Esta interpretación se alinea con las nuevas corrientes como la de Durana et al. (2025) que abogan por asociar la predicción con una acción específica y unos costos de error, ya que el modelo será tan bueno como el uso que se haga de él (revisión, bloqueo, negociación, cambios en la política interna).

Relevancia de las variables posibilita determinar qué predictores respaldan la clasificación, el Random forest con tres variables (peso_unitario, costo_logístico_ratio y precio_unitario) engloban el 75.6% de la importancia. la señal se enfoca en las métricas por unidad y en el peso proporcional de los costos relacionados con la logística. La empresa no puede traspasar el costo a las ventas, un incremento en la relación de costos logísticos disminuye el margen operativo, particularmente en compras. valores extremos de precio_unitario pueden estar relacionados con cambios en la especificación, variaciones de calidad, equivocaciones al digitalizar o condiciones

comerciales no habituales. Las variables mencionadas son útiles como disparadores de control, ya que un peso_unitario elevado aumenta la exposición a los fletes y provoca sensibilidad frente a modificaciones de tarifas o recargas. Esta atención a variables intensivas por unidad es consistente con la aplicación de indicadores derivados que codifican relaciones no lineales en la calificación de riesgo, tal como lo describen Hamdi et al. (2024).

Comparar modelos, exige validación externa, además de métricas, coherentes con la meta fijada. Papík y Papíková (2025) señalan que el rendimiento varía, dependiendo del método de evaluación y las técnicas de remuestreo; por eso, lo mejor es separar entrenamiento y prueba o usar validación cruzada. En este estudio, la evaluación en el conjunto total actúa como un punto de partida, y aconsejo emplear la partición train/test, para estimar el desempeño previsto en datos novedosos.

Finalmente, para reforzar la hacia la discusión la prevención, vale la pena enlazar las salidas del modelo con acciones concretas. Nguyen et al. (2024) plantean el uso de explicaciones contrafactuales en predicción de angustia para inferir los cambios mínimos que disminuyan el riesgo estimado, complementando la interpretabilidad de árboles y variables de importancia en soluciones accionables. Una salida así se fusiona fácilmente con una calculadora en Python, permitiendo al analista ingresar detalles de compra para ver clase y probabilidad. En la operativa, esto resulta en decisiones sencillas; como revisar documentos, renegociar el flete, cambiar la cantidad o incluso un proveedor solo si la probabilidad excede un límite interno.

CONCLUSIONES

La revisión de literatura apoyó que el aprendizaje automático se aplica con frecuencia para clasificar el riesgo de bancarrota o angustia, al incorporar diversas variables y capturar relaciones no lineales en los datos financieros. Los modelos basados en árboles y conjuntos son buenos sustitutos cuando se busca capacidad predictiva y lectura explicativa de variables, una condición favorable para estimar el riesgo en empresa importadora FEGOCAR.

Se entrenó un modelo supervisado de clasificación con Árbol de Decisión y Random Forest sobre datos de transacciones 2021–2025 (n = 4.531), realizando limpieza de importaciones, codificación de variables categóricas y creación de métricas por unidad (precio_unitario, peso_unitario) y razón logística (costo_logistico_ratio). El entrenamiento logró predicciones confiables para estimar probabilidad y clase de riesgo, con estructura replicable usando la función calculadora para nuevas transacciones.

Los resultados mostraron una alta exposición operativa: KPI de riesgo 48,78% (2.210 transacciones de alto riesgo). Random Forest mejoró en la capacidad de detección de casos de riesgo, disminuyendo los falsos negativos en comparación con el árbol de decisión y manteniendo un buen control de falsos positivos. Las variables que más influyeron fueron peso_unitario, costo_logistico_ratio y precio_unitario, las que pueden guiar controles preventivos a operaciones con alta carga logística, precios unitarios extremos y patrones por proveedor de mayor gasto agregado.

RECOMENDACIONES

Para argumentar de manera más precisa la elección del modelo y el sistema de evaluación, es necesario actualizar el marco teórico con una matriz comparativa de investigaciones realizadas entre 2021 y 2025 que traten sobre la predicción de quiebra/distress utilizando árboles y conjuntos. En esta matriz se deben incluir las variables empleadas, las dimensiones de la muestra, los métodos de validación y las métricas (AUC, F1 y retirada).

Para informar sobre el rendimiento fuera del entrenamiento y disminuir el ciclo de la evaluación interna, se sugiere implementar una división de datos en train/test o validación cruzada por año (como entrenar entre 2021 y 2024 y probar en 2025) y recalcular las métricas utilizando f1_score, souvenir_score, precision_score y AUC con predict_proba.

Crear una variable objetivo financiero por período para FEGOCAR (por ejemplo, estrés de liquidez, mora, margen negativo recurrente, deterioro de capital de trabajo) y entrenar el modelo con esa etiqueta, para aproximar la predicción a riesgo financiero real y no sólo a riesgo operativo transaccional.

Implementar una política de control basada en la salida del modelo: alertas por transacción con límites de probabilidad, revisión prioritaria de proveedores conforme a la tasade riesgo y al gasto total, así como el monitoreo mensual del KPI de riesgo, manteniendo un registro de las medidas correctivas y su verificación posterior en términos de costos o devoluciones.

Para FEGOCAR, desarrollar una variable objetivo financiero por período (como el estrés de liquidez, los pagos atrasados, un margen negativo sostenido o un capital de trabajo deteriorado) y entrenar el modelo con esa etiqueta para que la aproximación del resultado refleje no solo el riesgo operativo transaccional, sino también el riesgo financiero real.

Evaluar modelos complementarios como LightGBM/XGBoost y calibración de probabilidades (isotónica o Platt) para que la probabilidad emitida pueda ser interpretada como riesgo estimado, y cotejar su rendimiento con el de Random Forest y Árbol de Decisión utilizando el mismo método de validación temporal.

Agregar variables de tiempo (como la estacionalidad, la tendencia mensual, la variación de fletes o las alteraciones anuales) y emplear métodos para detectar la deriva o para realizar reentrenamiento programado, con el objetivo de preservar el desempeño del modelo cuando se producen cambios en los precios logísticos, en la mezcla de proveedores o en las condiciones de importación.

REFERENCIAS

Adan Gallo, J., Munar López, L., Romero Duque, G., & Gordillo Galeano, A. (2022). Nuevos desafíos de las pequeñas y medianas empresas en tiempos de pandemia. *Revista Tecnura, 26*(72), 185-208. https://doi.org/10.14483/22487638.17879

Ahmed Salman, H., Kalakech, A., & Steiti, A. (2024). Descripción general del algoritmo de bosque aleatorio. *Revista Babilónica de Aprendizaje Automático*, 69-79. https://doi.org/10.58496/BJML/2024/007

Almadani, B., Kaisar, H., Rashid Thoker, I., & Aliyu, F. (2025). Un estudio sistemático de los sistemas distribuidos de apoyo a la toma de decisiones en el ámbito sanitario. *Sistemas , 13*(3), 57. https://doi.org/10.3390/systems13030157

Altamirano Pérez, H. R. (Febrero de 2025). *El Impacto de la Inteligencia Artificial y el Machine Learning en la valoracion de activos finacieros.* www.instituto-ohiggins.edu.ec: https://www.instituto-ohiggins.edu.ec/wp-content/uploads/2025/02/El-Impacto-de-la-Inteligencia-Artificial-y-el-Machine-Learning-en-la-Valoracion-de-Activos-Financieros.pdf

Andino Chancay, T., Rodríguez López, V., Párraga Mogrovejo , M., & Molina Quiroz, C. (2022). Pequeñas y medianas empresas y la política comercial internacional del Ecuador. *Revista de Ciencias Sociales (Ve), XXVIII*(4), 448-469. https://www.redalyc.org/journal/280/28073811028/html/?utm_source

Arno, H., Mulier, K., Baeck, B., & Demeester, T. (2025). Predicción del fracaso empresarial a partir de datos textuales y tabulares con interpretaciones a nivel de oración. *Revista Ann Oper Res , 353* , 667–692. https://doi.org/10.1007/s10479-025-06574-z

Balcan, M.-F., & Sharma, D. (2025). Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25) Sister Conferences Best Papers Track. *Learning Accurate and Interpretable Decision Trees (Extended Abstract)∗.* https://www.ijcai.org/proceedings/2025/1205.pdf

Banco Bilbao Vizcaya Argentaria- BBVA. (15 de Julio de 2024). *'Machine learning': ¿qué es y cómo funciona el maestro en reconocer patrones?* www.bbva.com: https://www.bbva.com/es/innovacion/machine-learning-que-es-y-como-funciona/

Banco Interamericano de Desarrollo - BID. (8 de Octubre de 2024). *BID, BID Invest y BID Lab lanzan su informe 2024 de resultados e impacto de operaciones* . El Banco Interamericano de Desarrollo: https://www.iadb.org/es/noticias/bid-bid-invest-y-bid-lab-lanzan-su-informe-2024-de-resultados-e-impacto-de-operaciones

Bonilla Sierra, I. (Julio de 2025). *Predicción de quiebras financieras utilizando machine learning* . www.repositorio.comillas.edu: https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/95097/TFM%20-%20Sierra%20Bonilla%2C%20Luis.pdf?sequence=1&isAllowed=y

Broby, D. (2022). El uso del análisis predictivo en finanzas. *Revista de finanzas y ciencia de datos, 8*, 145-161. https://doi.org/10.1016/j.jfds.2022.05.003

Cámara Maritima del Ecuador - CAMAE. (7 de Diciembre de 2021). *Movimiento de cargas de importación en los puertos Ecuatorianos*. Cámara Marítima del Ecuador: https://www.camae.org/autoridad-portuaria-de-guayaquil/puerto-de-guayaquil-recibe-el-mayor-porcentaje-de-importaciones-maritimas/#:~:text=%C2%BFLe%20gust%C3%B3%20el%20art%C3%ADculo?&text=Seg%C3%BAn%20datos%20del%20Banco%20Central,de%20ingreso%20terrestre%

Cargua Pilco, E. (Agosto de 2020). *LA EDUCACIÓN FINANCIERA COMO BASE DE DESARROLLO PARA LOS COMERCIANTES DE LA EMPRESA PÚBLICA MUNICIPAL MERCADO DE PRODUCTORES AGRÍCOLAS SAN PEDRO DE RIOBAMBA.* Escuela Superior Politécnica del Chimborazo : https://dspace.espoch.edu.ec:8080/server/api/core/bitstreams/593bd7c9-d713-45b0-b3f6-cf7b42b39c12/content

Chacón Marín, S., Correa Jaramillo, J., Vasquez Merchan, I. L., & Balzan, A. (Julio de 2024). *Tendencias en Mercados Emergentes.* Sociedad de Doctores e Investigadores de Colombia: https://www.researchgate.net/publication/383232384_Tendencias_en_Mercados_Emergentes

Comex. (2022). *Informe sobre comercio exterior y uso de analítica de datos en empresas importadoras*. Comité Empresarial Ecuatoriano: https://www.produccion.gob.ec/comex/

Comisión Económica para América Latina y el Caribe-CEPAL. (2025). *Acerca de Microempresas y Pymes*. Comisión Económica para América Latina y el Caribe: https://www.cepal.org/es/temas/micro-pequenas-medianas-empresas-mipyme/acerca-microempresas-pymes

Constitución de la República del Ecuador.-CRE. (25 de Enero de 2021). *CONSTITUCIÓN DE LA REPÚBLICA DEL ECUADOR*. CONSTITUCIÓN DE LA REPÚBLICA DEL ECUADOR 2008: https://www.defensa.gob.ec/wp-content/uploads/downloads/2021/02/Constitucion-de-la-Republica-del-Ecuador_act_ene-2021.pdf

Corte Nacional de Justicia . (12 de Mayo de 2021). *ACUMULACIÓN DE PROCESOS EN MATERIA CONCURSAL.* Corte Nacional de Justicia : https://www.cortenacional.gob.ec/cnj/images/pdf/consultas_absueltas/No_Penales/Procesal/171.pdf

Dasilas, A., & Rigani, A. (2024). Técnicas de aprendizaje automático en la predicción de quiebras: una revisión sistemática de la literatura. *Revista ELSEVIER, 255*. https://doi.org/10.1016/j.eswa.2024.124761

Datascientest. 2025. *Machine Learning: definición, funcionamiento, usos*. www.datascientest.com: https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

Dunkley Hickin, K. (28 de Octubre de 2021). *Una introducción a la teoría del árbol de decisiones*. Análisis de Precisión : https://www.precision-analytics.ca/articles/an-introduction-to-decision-tree-theory/

Erazo Peña, J. (Agosto de 2024). *Evaluación de la eficiencia en la detección de tráfico malicioso.* www.dspace.espoch.edu.ec: https://dspace.espoch.edu.ec:8080/server/api/core/bitstreams/ca8013ed-b796-4691-9f4a-075c51bce402/content

Espinosa Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Revista Ingeniería, investigación y tecnología, 21* (3). https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Fernández Solís, M., Fernández Ronquillo, M., & Lasso Davila, W. (2024). Administración del Riesgo Financiero en las Pymes de Ecuador. *Revista Multidisciplinar Veritas , 5*(3), 333–355. https://doi.org/10.61616/rvdc.v5i3.207

Franco Gómez, Y. A. (4 de Juli de 2023). *Un análisis bibliométrico de la predicción de quiebra empresarial con Machine Learning*. Universidad Externado de Colombia : https://bdigital.uexternado.edu.co/entities/publication/d6d7ccab-1eb8-4687-91e9-7d33b17291c6

Gajdosikova, D., & Valaskova, K. (2023). Bankruptcy prediction model development and its implications on financial performance in Slovakia.

Sciendo. Economics and Culture, 20(1), 30-42. https://doi.org/10.2478/jec-2023-0003

Gobierno de España. (19 de Abril de 2023). *Qué es la Inteligencia Artificial*. Plan de Recuperación, Transformación y Resiliencia : https://planderecuperacion.gob.es/noticias/que-es-inteligencia-artificial-ia-prtr

Guamán Briones, T. G., & Goya Contreras, R. E. (2025). Análisis financiero de pequeñas y medianas empresas de Servicios del sector Urdesa Central del norte de Guayaquil, periodo 2023-2024. *Revista Científica Zambos, 4*(1), 249-272. https://doi.org/10.69484/rcz/v4/n1/89

Hernández, R., Fernández-Collado, C., & Baptista, P. (2006). *Metodología de la investigación* (4° ed.). México: McGraw-Hill. http://187.191.86.244/rceis/registro/Metodolog%C3%ADa%20de%20la%20Investigaci%C3%B3n%20SAMPIERI.pdf

Herrera Sánchez, M. J., & Casanova Villalba, C. I. (2024). Inteligencia artificial y su impacto en la transformación de la gestión financiera. *Space Scientific Journal of Multidisciplinary , 2*(1), 52-64. https://doi.org/10.63618/omd/ssjm/v2/n1/43

IBM. Cognos Analytics (17 de Abril de 2025). *Árbol de clasificación*. IBM Cognos Analytics: https://www.ibm.com/docs/es/cognos-analytics/12.1.x?topic=tests-classification-tree

Imani, M., Beikmohammadi, A., & Arabnia, H. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies, 13*(3), 88. https://doi.org/10.3390/technologies13030088

Instituto Nacional de Estadística y Censos - INEC. (2021). *Directorio de Empresas y Establecimientos 2020*. Instituto Nacional de Estadisticas y Censos: https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Economicas/DirectorioEmpresas/Directorio_Empresas_2020/Principales_Resultados_DIEE_2020.pdf

Isaac, D., & Caicedo, A. (2023). Relación entre los indicadores financieros del modelo Altman Z y el puntaje Z. *RETOS Revista de Ciencia de Administración y Economía, 13*(25). https://doi.org/10.17163/ret.n25.2023.09

ISO. (2023). *Gestión de riesgos: directrices*. ISO 31000:2018: https://www.iso.org/standard/65694.html

Jardón Gallegos, M., Pérez Zúñiga, E., Sánchez Herrera, J., & Guallpa Calva, M. (06 de Agosto de 2024). *Análisis del riesgo de insolvencia en las*

empresas. www.reincisol.com: file:///C:/Users/Usuario/Downloads/Dialnet-AnalisisDelRiesgoDeInsolvenciaEnLasEmpresasManufac-10004620.pdf

Jones, S. (2023). Una revisión de la literatura sobre modelos de predicción de fracasos corporativos. *Journal of Accounting Literature , Emerald Group Publishing Limited, 45*(2), 364-405. https://doi.org/10.1108/JAL-08-2022-0086

Koukaras, P., & Tjortjis, C. (2025). Preprocesamiento de datos e ingeniería de características para minería de datos: técnicas, herramientas y mejores prácticas. *Revista AI, 6*(10), 257. https://doi.org/10.3390/ai6100257

Laguillo Diaz, G. (Noviembre de 2020). *Prediccion de insolvencias en el sector economico. Un analisis comprativo.* www.scholar.google.com: https://dialnet.unirioja.es/servlet/tesis?codigo=76910

Ley Orgánica de Protección de Datos Personales - LOPDP. (26 de Mayo de 2021). *LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES* . LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES : https://www.finanzaspopulares.gob.ec/wp-content/uploads/2021/07/ley_organica_de_proteccion_de_datos_personales.pdf

Leyva Ferreiro, G., & Gómez García, S. (09 de Diciembre de 2021). *Utilidad de los modelos de predicción de fracaso y su aplicabilidad en las cooperativas.* www.scielo.sld.cu: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2073-60612019000300013

Ludeña Dávila, M. K., & Tonon Ordóñez, L. B. (2021). Calculando el riesgo de insolvencia, de los métodos tradicionales a las redes neuronales artificiales. Una revisión de literatura. *INNOVA Research Journal , 6*(3), 270-287. https://doi.org/10.33890/innova.v6.n3.2021.1790

Luna Rizo, M., Daza Ramírez, M. T., & Lozoya Arandia, J. (2024). *Tendencias de la Inteligencia Artificial en Educación .* Universidad de Guadalajara : https://mta.udg.mx/sites/default/files/adjuntos/tendencias_de_la_inteligencia_artificial.pdf

Malakauskas, A., & Lakštutienė, A. (2021). Predicción de dificultades financieras para pequeñas y medianas empresas mediante técnicas de aprendizaje automático. *Economía de la ingeniería , 32*(1), 4-14. https://doi.org/10.5755/j01.ee.32.1.27382

Marquez, A. (18 de Octubre de 2022). *Caso 5. Bosques aleatorios random forest con datos advertising en Python*. RPubs : https://rpubs.com/alexmaar/957877

Martins, J. (20 de Febrero de 2025). *Qué es la gestión de riesgos y cómo aplicarla a tu proyecto en solo 6 pasos*. Asana: https://asana.com/es/resources/project-risk-management-process

McCartney, A. (16 de Octubre de 2023). *Las 10 principales tendencias tecnológicas estratégicas de Gartner para 2024*. Gartner: https://www.gartner.es/es/articulos/las-10-principales-tendencias-tecnologicas-estrategicas-de-gartner-para-2024

Medina Merino, R. F., & Ñique Chacón, C. I. (2020). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases , 10*. https://doi.org/10.26439/interfases2017.n10.1775

Mendoza, R. (2006). *Investigación cualitativa y cuantitativa. Diferencias y limitaciones.* Piura. https://www.insp.mx/resources/images/stories/Centros/nucleo/docs/dip_lsp/investigacion.pdf

Ministerio de Producción . (Enero de 2025). *Boletín de cifras del Sector Productivo* . https://www.produccion.gob.ec/wp-content/uploads/2025/01/Boletin-de-Produccion-ENE2025.pdf

Molina Panchi, P., & Molina Panchi, D. (2024). Modelos de Predicción de Fragilidad Empresarial: Una Herramienta para Detectar la Bancarrota. *Revista de investigación SIGMA, 11*(1). https://doi.org/10.24133/20hwq783

Molina Panchi, P., Flores Cevallos, K., Flores Tapia, C., & Molina Panchi, D. (2023). Modelo de predicción de quiebra en empresas de comercio en Ecuador: Uso del modelo logístico de Ohlson. *Revista Científica Ecociencia , 10*(3). https://doi.org/10.21855/ecociencia.103.812

Moloina Ayala, C., & Piedra Aguilera, M. (01 de Febrero de 2024). *Estudio del impacto de estrategias financieras en la sostenibilidad y productividad de las PYMES.* www.revistas.uazuay.edu.ec/: https://revistas.uazuay.edu.ec/index.php/udaakadem/article/view/832/1272

Monelos , p., Piñerio , C., & Rodrigez , M. (Diciembre de 2021). *Predicción del fracaso empresarial. Una contribución a la síntesis de prediccion.* www.scielo.cl: https://www.scielo.cl/pdf/ede/v43n2/art01.pdf

Ocampo Alvarado, A. M. (2024). Efectos de la transformación digital en el sector contable y financiero en Ecuador. *Revista Academo, 11*(3), 233-241. https://doi.org/10.30545/academo.2024.set-dic.2

Olavsrud, T. (30 de Enero de 2024). *Sistemas de apoyo a la toma de decisiones: la herramienta clave para el análisis de datos empresariales.* CIO: https://www.cio.com/article/1314134/sistemas-de-apoyo-a-la-toma-de-decisiones-la-herramienta-clave-para-el-analisis-de-datos-empresariales.html

Organización de las Naciones Unidas- ONU. (2024). *Informe comercio y el desarrollo* . CONFERENCIA DE LAS NACIONES UNIDAS SOBRE COMERCIO Y DESARROLLO: https://unctad.org/system/files/official-document/tdr2023_es.pdf

Osinaga Flores, L. C. (2024). Fortalezas y Debilidades Financieras de las Pymes de Servicios de Santa Cruz, Bolivia. *Revista Economía y Negocios , 15*(1). https://www.redalyc.org/journal/6955/695578766010/

Pardeshi, B. (2022). Análisis de regresión logística para la predicción de fracasos financieros: evidencia de empresas del sector público central en India. *Vision: The Journal of Business Perspective*. https://doi.org/10.1177/097226292211352

Pérez Prieto, M. E., Bravo Armijos, Á. J., & Sánchez González, W. G. (2025). Big Data Analytics en Instituciones Financieras: Potenciando la Toma de Decisiones y la Eficiencia Operativa. *Polo del Conocimiento , 10*(11), 1076-1099. https://doi.org/10.23857/pc.v10i11.10685

Pérez, Y. (18 de Septiembre de 2025). *Economías de mercados emergentes: definición, crecimiento y actores clave*. Investopedia: https://www.investopedia.com/terms/e/emergingmarketeconomy.asp

Plaza, D. (20 de Julio de 2024). *Economía Computacional y el Uso de Machine Learning en la Predicción de.* www.remulcio.com: file:///C:/Users/Usuario/Downloads/ART_21_18-33%20(1).pdf

Riquelme Benítez, C. R., & Pereira Benítez, M. A. (2024). Inteligencia artificial como herramienta de organización de actividades profesionales y personales. *Revista Científica Ciencias Sociales*. https://doi.org/10.53732/rccsociales/e601502

Rivadeneira, J., Saltos, R., Rivera, M., & Carpio, R. (2022). Predicción de la quiebra empresarial en el sector agroindustrial de la ciudad de Machala. *Revista ACI Avances en Ciencias e Ingenierías, 14*(2). https://doi.org/10.18272/aci.v14i2.2695

Romero Espinosa, F. (2023). *Alcances y limitaciones de los modelos de capacidad predictiva en el análisis del fracaso empresarial*. ResearchGate: https://www.researchgate.net/publication/268280733_Alcances_y_limitaciones_de_los_modelos_de_capacidad_predictiva_en_el_analisis_del_fracaso_empresarial

Ruiz Díaz de Salvioni, V. V., & Armoa, A. (2023). La importancia de la Minería de Datos como una herramienta estratégica en las Empresas. *Ciencia Latina Revista Científica Multidisciplinar,* *7*(1), 9267-9276. https://doi.org/10.37811/cl_rcm.v7i1.5119

Salazar Xirinachs, J. M. (29 de Marzo de 2023). *En 2023 el crecimiento será más lento en América Latina y el Caribe: así es como se puede revertir el ciclo*. CEPAL: https://www.cepal.org/es/articulos/2023-2023-crecimiento-sera-mas-lento-america-latina-caribe-asi-es-como-se-puede-revertir

Scherge, V., Terceño, A., & Vigier, H. (04 de Abril de 2018). *Revisión crítica de los modelos de predicción.* www.scholar.google.com: file:///C:/Users/Usuario/Downloads/admin,+RAYO-40-153-180.pdf

Scikit Learn. (2025). *Árboles de decisión* . Scikit Learn: https://scikit-learn.org/stable/modules/tree.html

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

Shmueli, G., & Koppius, O. (7 de Agosto de 2025). Análisis predictivo en la investigación de sistemas de información. *Revista Electrónica SSRN, 35*(3), 553-572. Revista Electrónica SSRN: https://www.researchgate.net/publication/220260303_Predictive_Analytics_in_Information_Systems_Research

Suescum Coelho, C.-E., Cisneros Laguna, M. V., & Suescum Coelho, C. (2025). Riesgos financieros en PYMES importadoras venezolanas: una perspectiva de gestión sostenible alineada a los ODS 8 y 9. *Star of Sciences Multidisciplinary Journal,* *2*(1), 1-13. https://doi.org/10.63969/arygmd93

Superintendencia de Protección de Datos Personales - SPDP. (12 de Mayo de 2025). *PLAN NACIONAL DE PROTECCIÓN DE DATOS PERSONALES* . Superintendencia de Protección de Datos Personales : https://spdp.gob.ec/plan-nacional-pdp/#

SuperCias. (18 de Junio de 2025). *Superintendencia destacó crecimiento exponencial de las S.A.S. en encuentro con el sector productivo*.

Superintendencia de Compañias, Valores y Seguros : https://www.supercias.gob.ec/portalscvs/Noticias/Noticias.php?seccion=noticia-18062025#:~:text=En%20cuanto%20a%20los%20sectores,herramientas%20jur%C3%ADdicas%20modernas%20y%20accesibles.

Terreno, D., Pérez, J., & Sattler, S. (2024). Un modelo jerárquico para la predicción de insolvencia empresarial. Aplicación de análisis discriminante y árboles de clasificación. *ResearchGate, 43*(91), 51-76. https://doi.org/10.15446/cuad.econ.v43n91.105115

Toapanta Cunalata, D. G., Ortiz Betancourt, W., & Natividad, B. G. (2024). Aplicación de la Inteligencia Artificial para la Detección de Riesgos Financieros: Un Estudio de Programación Computacional. *Revista de Investigación Sigma, 11*(1). https://doi.org/10.24133/fm72c767

Trujillo Coloma, M. J., Guerrero Zambrano, E. O., Moreno Díaz, V. H., & Tejada Castro, M. I. (2025). Modelado predictivo del riesgo financiero en empresas de tecnología mediante una plataforma de inteligencia de negocio. *Revista RECIAMUC, 9*(4), 89-103. https://doi.org/10.26820/reciamuc/9.(4).diciembre.2025.89-103

UNIR. (6 de Mayo de 2025). *¿Qué es el machine learning o aprendizaje automático y cómo funciona?* Universidad Internacional de la Rioja: https://mexico.unir.net/noticias/ingenieria/machine-learning/

Universidad Internacional de la Rioja. (30 de Noviembre de 2023). *¿Qué es un analista financiero y qué hace?* Unir Revista: https://www.unir.net/revista/empresa/analista-financiero/

Universidad Técnica Particular de Loja - UTPL. (19 de Junio de 2024). *Las pymes promueven la innovación tecnológica y sostenibilidad en el sector empresarial*. Universidad Técnica Particular de Loja: https://noticias.utpl.edu.ec/las-pymes-promueven-la-innovacion-tecnologica-y-sostenibilidad-en-el-sector-empresarial

Uriarte del Águila, C. (2024). Aplicación de la minería de datos como herramienta estratégica en la gestión del conocimiento. *Diginomics, 3*(123), 1-12. https://doi.org/10.56294/digi2024123

Uzcátegui, C., Zaldumbide, D., & Leite, E. (2024). *Revolución de la Investigación de Mercados Impulsada por la IA .* Editora Artemis . https://doi.org/10.37572/EdArt_081124352

Vaca Sigüeza, A. J., & Orellana Osorio, I. (23 de Junio de 2020). *Análisis de riesgo financiero en el sector de fabricación de otros productos minerales no metálicos del Ecuador.* https://doi.org/10.25097/rep.n32.2020.05

Valente Galán, M. M. (29 de Agosto de 2024). *Propuesta de un plan de gestión de RSC para las MIPYMES del sector manufacturero en la ciudad de Guayaquil*. Universidad Católica Santiago de Guayaquil: http://repositorio.ucsg.edu.ec/bitstream/3317/23314/1/UCSG-C339-22923.pdf

Vargas Charpentier, J. a. (Marzo de 2020). *Modelos de breaver, Ohlosn y altman Son realmente capaces de pr4edecir la banca rota del sector empresarial.* www.Dialnet.unirioja.es.

Vargas Ochoa, M. O. (30 de Abril de 2020). *Máquinas de Soporte Vectorial Support Vector Machine SVM en regresión*. RPubs : https://rpubs.com/movo/607465

Vilema Lara, P. H. (2022). *"DETECCIÓN DE FALLOS EN MANTENIMIENTO PREDICTIVO UTILIZANDO EL MÉTODO DE APRENDIZAJE DE MÁQUINA RANDOM FOREST"*. Escuela Superior Politécnica de Chimborazo: https://dspace.espoch.edu.ec:8080/server/api/core/bitstreams/74d0feca-ee97-4d61-a090-c7f80df373f8/content

Villegas Rojas, J. S., & Altamirano Freire, J. L. (2025). Transformación Digital y Modelos de Negocio. *Polo del Conocimiento , 10*(1), 2899-2919. https://doi.org/10.23857/pc.v10i1.8871

Yance Carvajal, C., Solís Granda, L., Burgos Villamar, I., & Hermida, L. (Junio de 2021). *LA IMPORTANCIA DE LAS PYMES EN EL ECUADOR*. Revista Observatorio de la Economía Latinoamericana, Ecuador: https://sga.unemi.edu.ec/media/evidenciasiv/2017/06/07/articulo_20176713520.pdf

Yizinet. (2022). *Ganancia de Información y Entropía* . www.yizinet.com: https://www.yizinet.com/ganancia-de-informacion-y-entropia/

Zambrano, R., San Andrés, P., & Paredes, I. (2019). Factores que inciden en las exportaciones de las PyMEs del Ecuador. Período 2012-2016. *Revista Espacios , 40*(40), 4ç. https://www.revistaespacios.com/a19v40n40/19404004.html

DECLARACIÓN Y AUTORIZACIÓN

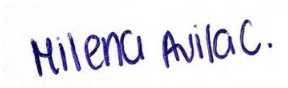
Nosotros, **Avila Coello, Milena Jamilet**, con C.C: # **0929740082** y **Mendoza Jumbo, Jair Daniel** C.C: # **0957561251** autor/a del trabajo de titulación: **Modelo de datos para predecir quiebras de pymes importadoras en Guayaquil: análisis con árboles de decisión y random forest,** previo a la obtención del título de **Licenciado en Negocios Internacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

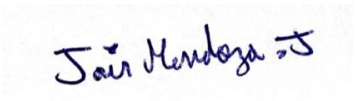
2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, **12 del mes de febrero del año 2026**

LOS AUTORES



f. _____
Avila Coello, Milena Jamilet
C.C: # 0929740082



f. _____
Mendoza Jumbo, Jair Daniel
C.C: # 0957561251

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA

FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN

TEMA Y SUBTEMA:	**Modelo de datos para predecir quiebras de pymes importadoras en Guayaquil: análisis con árboles de decisión y random forest.**		
AUTOR(ES)	**Avila Coello, Milena Jamilet** **Mendoza Jumbo, Jair Daniel**		
REVISOR(ES)/TUTOR(ES)	**Carrera Buri, Félix Miguel**		
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil		
FACULTAD:	**Economía y Empresa**		
CARRERA:	**Negocios Internacionales**		
TITULO OBTENIDO:	**Licenciados en Negocios Internacionales**		
FECHA DE PUBLICACIÓN:	**12** de **febrero** de **2026**	**No. DE PÁGINAS:**	119 p.
ÁREAS TEMÁTICAS:	**Comercio internacional, Auditoría de gestión, Auditoría financiera, Gestión de riesgos.**		
PALABRAS CLAVES/ KEYWORDS:	**Random Forest, Árbol de Decisiones, predecir quiebras, Phyton, economía.**		

RESUMEN/ABSTRACT (150-250 palabras):
Esta investigación evalúa el riesgo de quiebra operativo vinculado a la presión financiera en una compañía importadora de Guayaquil (FEGOCAR) por medio de aprendizaje automático, utilizando unas herramientas de alerta para el control interno. Se entrenó un clasificador supervisado con 4531 transacciones entre 2021 y 2025 empleando variables económicas y operativas, por ejemplo precio transporte seguro cantidad peso neto país de origen el tipo de mercancía modelo y además proveedor. Se llevó a cabo la limpieza de datos, la conversión del formato monetario a un estándar, la codificación one-hot de las variables categóricas y el desarrollo de métricas derivadas: precio_unitario, peso_unitario y costo_logistico_ratio, siendo que para clasificar operaciones como de alto o bajo riesgo, se entrenaron un Random Forest y un Árbol de Decisión con equilibrio entre las clases. El KPI de riesgo operativo se ubicó en 48,78% (2.210 transacciones en alto riesgo). Forest logró AUC 0,824 y exactitud 0,827; el Árbol de Decisión logró AUC 0,817 y exactitud 0,819. La importancia de variables dio como predictores de mayor contribución a peso_unitario, costo_logistico_ratio y precio_unitario. Se generaron rankings por proveedor: tasa de riesgo y gasto total, así como reglas legibles del árbol y una calculadora para estimar el riesgo en nuevas transacciones, útil para auditoría focalizada, validación de costos logísticos y control de transacciones inusuales.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:** +593-4- **(registrar teléfonos)**	E-mail: milena.avila@cu.ucsg.edu.ec Jair.mendoza@cu.ucsg.edu.ec
CONTACTO CON LA INSTITUCIÓN (C00RDINADOR DEL PROCESO UTE)::	**Nombre: Freire Quintero Cesar Enrique**	
	Teléfono: +593 990090702	
	E-mail: cesar.freire@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA

Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	