Escuela Superior Politécnica del Litoral

PROYECTO DE TITULACIÓN:

Análisis Predictivo de Patrones en Transacciones Comerciales en una empresa exportadora del sector acuícola utilizando Machine Learning para la Auditoría Financiera Moderna

Previa la obtención del Título de:

MAGÍSTER EN CONTABILIDAD Y AUDITORÍA

MENCIÓN ANALÍTICA DE DATOS

Presentado por:

MIRANDA GÓMEZ KEVIN ANDRÉS

Guayaquil – Ecuador

2025

AGRADECIMIENTO

Deseo expresar mi más profundo y sincero agradecimiento a quienes hicieron posible la culminación de este proyecto de titulación. En primer lugar, al MsC. César Olmedo Navarro y al MsC. Benigno Alfredo Armijos de la Cruz, tutores de esta investigación, quienes, con paciencia, compromiso y un acompañamiento constante me brindaron la orientación académica necesaria para llevar a término este trabajo. Sus observaciones, recomendaciones y valioso feedback enriquecieron cada etapa del proceso, permitiéndome superar las dificultades que se presentaron en el camino y alcanzar los objetivos planteados.

Extiendo también mi gratitud a la MsC. Caterine Narcisa Vásquez Castro, coordinadora académica de la maestría de contabilidad y auditoría, por su apoyo permanente, su comprensión ante las circunstancias que retrasaron la entrega del proyecto y su capacidad de gestión que me permitió continuar hasta concluir con éxito esta etapa académica. Su empatía y profesionalismo fueron determinantes para que este esfuerzo pudiera consolidarse.

A todos ellos, expreso mi más sincero reconocimiento por la confianza depositada en mí, por su orientación y por el acompañamiento humano y académico que hicieron posible que este proyecto se concrete y que hoy pueda dar un paso más en mi formación profesional.

COMITÉ DE EVALUACIÓN

Firmado electrónicamente por:
CESAR ALEJANDRO OLMEDO NAVARRO
Validar únicamente con FirmaEC

MBI. César Olmedo Navarro

Tutor del Proyecto

Firmado electrónicamente por:
BENIGNO ALFREDO ARMIJOS DE LA CRUZ
Validar únicamente con FirmaEC

MSc. Alfredo Armijos de la Cruz

Evaluador 1

CATERINE NARCISA VASQUEZ CASTRO

Firmado digitalmente por CATERINE NARCISA VASQUEZ CASTRO
Fecha: 2025.09.12 11:10:53 -05'00'

MSc. Caterine Vásquez Castro

Presidenta

Declaración Expresa

Yo Miranda Gómez Kevin Andrés acuerdo y reconozco que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me/nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi/nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique al autor que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 29 de Agosto del 2025.

Kevin Andrés Miranda Gómez

COMPROMISO DE AUTOR

Yo, Kevin Andrés Miranda Gómez**, declaro que:**

El contenido del presente documento es original y constituye un reflejo de mi trabajo personal. Manifiesto que, ante cualquier notificación de plagio, autoplagio, copia o falta a la fuente original, soy responsable directo legal, económico y administrativo sin afectar al Director del trabajo, a la Universidad y a cuantas instituciones hayan colaborado en dicho trabajo, asumiendo las consecuencias derivadas de tales prácticas.

Firma

AUTORIZACIÓN PUBLICACIÓN ELECTRÓNICA

Guayaquil, 29 de agosto del 2025

Por este medio autorizo la publicación electrónica de la versión aprobada de mi Proyecto Final bajo el título "Análisis Predictivo de Patrones en Transacciones Comerciales en una empresa exportadora del sector acuícola utilizando Machine Learning para la Auditoría Financiera Moderna" en el campus virtual y en otros espacios de divulgación electrónica de esta Institución.

Informo los datos para la descripción del trabajo:

Título	Análisis Predictivo de Patrones en Transacciones Comerciales en una empresa exportadora del sector acuícola utilizando Machine Learning para la Auditoría Financiera Moderna
Autor	Kevin Andrés Miranda Gómez
Resumen	Este proyecto se orienta a resolver la necesidad de detectar de manera anticipadas patrones inusuales en las transacciones comerciales de una empresa exportadora del sector acuícola, con el propósito de fortalecer la auditoría financiera moderna y contemporánea y minimizar riesgos asociados a prácticas irregulares o fraudes.
Programa	Maestría en Contabilidad y Auditoría Mención Analítica de Datos
Palabras clave	Machine Learning, Auditoría, Inteligencia Artificial
Contacto	kevmiran@espol.edu.ec; kamg95@hotmail.com

Atentamente,

Firma

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Este proyecto se orienta a resolver la necesidad de detectar de manera anticipadas patrones inusuales en las transacciones comerciales de una empresa exportadora del sector acuícola, con el propósito de fortalecer la auditoría financiera moderna y contemporánea y minimizar riesgos asociados a prácticas irregulares o fraudes. El objetivo principal del trabajo fue utilizar modelos de análisis predictivo que, mediante la aplicación de técnicas de aprendizaje automático, posibilite la identificación de comportamiento extraños y tendencias relevantes en los mercados de exportación y mercado local. La metodología adoptada se basó en el modelo CRISP-DM, que se caracteriza en los proyectos de ciencia de datos, integrando las fases de comprensión del negocio y de los datos, preparación, modelado, evaluación y diseño de la propuesta; se clasifica como un proyecto de solución tecnológica. La estructura del trabajo empieza desde la exposición del problema hasta la implementación e interpretación de algoritmos de machine learning supervisados y no supervisados, que incluyen DBSCAN, isolation forest, random forest y regresión logística, para finalmente formular una propuesta de aplicación. Como parte significativa, el proyecto permitió identificar con precisión patrones y anomalías que optimizan el control interno en contabilidad, respaldan decisiones estratégicas y dotan a la empresa de una herramienta muy útil en la nube para el monitoreo continuo de sus transacciones, incorporando capacidades analíticas de un nivel alto a su sistema de control interno.

INTRODUCCIÓN

La presente investigación contiene el análisis predictivo de patrones en transacciones comerciales de una empresa exportadora del sector acuícola en Ecuador, con el fin de fortalecer los procesos de auditoría financiera y control interno mediante la incorporación de técnicas de aprendizaje automático. La importancia de esta temática radica tanto en lo teórico como en lo práctico, por un lado se contribuye a la literatura especializada en el uso de inteligencia artificial aplicada a la detección de anomalías financieras; y por otro, ofrece a la empresa un recurso tecnológico de gran utilidad para prevenir fraudes, optimizar la gestión de riesgos y mejorar la toma de decisiones estratégicas. El problema identificado surge de la complejidad y el creciente volumen de las operaciones comerciales, lo que dificulta el control manual y puede dar lugar a inconsistencias o transacciones sospechosas que no son detectadas oportunamente, lo cual puede afectar a la transparencia de la organización y su rentabilidad.

El trabajo se clasifica como un proyecto de solución tecnológica, que va de la mano con la aplicación del modelo CRISP-DM, que integra fases de comprensión, preparación, modelado, evaluación y despliegue. El contexto de desarrollo corresponde a una empresa exportadora del sector acuícola que opera en mercados internacionales y locales, que enfrenta el reto de analizar grandes volúmenes de datos heterogéneos para garantizar la fiabilidad de la información. El objetivo principal de la propuesta es implementar un sistema de análisis predictivo capaz de identificar de forma temprana patrones atípicos y comportamiento anómalos, proporcionando una herramienta automatizada que complemente el trabajo de auditoría y refuerce el control de las operaciones.

El documento se estructura en varios capítulos. El capítulo uno se plantea el problema, se justifica el tema y se establecen los objetivos generales y específicos. El capítulo dos describe el

marco teórico y conceptual sobre auditoría financiera moderna, machine learning y técnicas de análisis de datos aplicadas a entornos comerciales. El capítulo tres desarrolla la metodología adoptada, especificando la aplicación del modelo CRISP-DM y los procedimientos de recolección y tratamiento de datos. El capítulo cuatro presenta el análisis exploratorio y la preparación de datos, detallando los procesos de limpieza, segmentación y detección inicial de patrones, se expone el modelado con algoritmos supervisados y no supervisados como DBSCAN, isolation forest, random forest y regresión logística, junto con la interpretación de resultados. En el capítulo cinco se formula la propuesta de aplicación práctica, conectando los hallazgos con la solución diseñada para optimizar los procesos de auditoría, además se recoge las conclusiones, resaltando la eficacia del modelo propuesto.

CAPÍTULO 1: PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción del problema o de la oportunidad

En la actualidad desde un contexto global, la auditoría financiera enfrenta desafíos sin precedentes debido al volumen, velocidad y complejidad de las transacciones comerciales, en este caso la industria camaronera, un sector crucial en la economía de muchos países costeros, no es ajena a esta problemática. La evidencia reciente en el comercio internacional de camarón, valorado en más de $22 mil millones anuales según la ((FAO), 2022), se caracteriza por su volatilidad y susceptibilidad a fluctuaciones de mercado, políticas comerciales, regulaciones sanitarias, cambios climáticos y enfermedades que afectan la producción.

Hasta el momento el problema central radica en la dificultad de los métodos tradicionales de auditoría para detectar eficazmente patrones anómalos, fraudes potenciales y riesgos financieros en tiempo real dentro de este sector bastante fluctuante.

La oportunidad surge en la aplicación del machine learning (en adelante denominado ML) a la auditoría financiera moderna para este giro de negocio. Sin embargo, la implementación efectiva de estas tecnologías enfrenta obstáculos significativos, como la falta de modelos específicos para la industria, la brecha de habilidades en los auditores, la resistencia al cambio y los desafíos éticos y regulatorios (Sun & Vasarhelyi, 2018).

La cuestión central en este proyecto es abordar estas problemáticas desarrollando un enfoque de machine learning para el análisis predictivo de patrones de transacciones en la industria camaronera, con el objetivo de mejorar la detección de anomalías y fraudes, adaptándose a las características únicas del sector, y también contribuir a la sostenibilidad del sector para reducir riesgos financieros y mejorar la transparencia operativa.

La auditoría financiera se apoya cada vez más en la automatización de procesos para monitorear en tiempo real grandes volúmenes de datos, pero en Ecuador muchas de estas empresas del sector camaronero aún utilizan sistemas manuales o semiautomáticos que imposibilitan el análisis predictivo eficiente de datos. Para (Mariano, Pedro, & José, 2021) En el área de contabilidad la falta de automatización genera un entorno a los que se deben recurrir a procedimientos tradicionales, pero enfocándolo en la auditoría sin un sistema adecuado de procesamiento de datos que para el contexto presente se centra en los datos comerciales, es imposible aprovechar las capacidades y beneficios del ML que requiere grandes volúmenes de datos estructurados y bien organizados para entrenar modelos predictivos y detectar patrones irregulares.

A pesar de estos obstáculos, las camaroneras ubicadas en Guayaquil tienen un gran potencial para beneficiarse de la modernización tecnológica. Según datos de la Cámara Nacional de Acuacultura ((CNA), 2024) esta ciudad concentra un porcentaje significativo de las empresas exportadoras de camarones, las cuales generaron un alto volumen de ingresos para el país en 2023. Con ingresos que superan los 5.4 mil millones de dólares a nivel nacional, el sector camaronero es uno de los más importantes de la economía ecuatoriana. La implementación de pronósticos predictivos en los procesos de auditoría podría ayudar a estas empresas a detectar irregularidades y optimizar sus controles internos. Sin embargo, para aprovechar al máximo estas tecnologías, es fundamental que las camaroneras inviertan en la automatización de sus sistemas contables y en la capacitación de su personal, garantizando una infraestructura adecuada para el análisis predictivo y la toma de decisiones basada en datos.

1.2. Justificación

El estudio aborda la intersección entre el aprendizaje automático y la auditoría financiera, esto surge gracias a la creciente digitalización, complejidad y volumen de transacciones por lo que se requiere evolucionar las técnicas tradicionales de auditoría, para incorporar enfoques basados en datos que permitan evaluar e identificar patrones, anomalías con mayor precisión. Para los autores (Vasarhelyi, Kogan, & Tuttle, 2015) el uso de machine learning puede mejorar e incluso optimizar la detección de fraudes y la calidad de informes financieros, al generar análisis automatizados y adaptativos que superan las limitaciones de las auditorías manuales.

Es por esto por lo que el proyecto promueve el desarrollo de enfoques predictivos que permitan anticiparse a los riesgos que se pueden generar en las auditorías, específicamente en este enfoque a una compañía perteneciente a la industria camaronera.

Este trabajo es pertinente porque utiliza un enfoque cuantitativo y predictivo, basado en la aplicación de algoritmos de machine learning, para el análisis de transacciones comerciales, la metodología es adecuada para abordar el problema de la detección de distintos patrones en grandes volúmenes de datos financieros, permitiendo evaluar la efectividad de modelos predictivos en la identificación de riesgos y fraudes.

Por lo tanto, la investigación es relevante para la industria camaronera, particularmente en empresas ecuatorianas de la provincia del guayas, que representan un porcentaje significativo de la producción y exportación del país, un sector caracterizado por depender de mercados internacionales y la variación de los precios para esto la aplicación de tecnologías avanzadas es vital para anticipar riesgos y detectar fraudes en tiempo real en pro de la sostenibilidad del negocio a largo plazo.

1.3. Objetivos

1.3.1 Objetivo general

Aplicar técnicas de Machine Learning para identificar patrones en transacciones comerciales con el fin de resolver de una mejor manera la toma de decisiones basada en datos y aumentando la transparencia en la auditoría financiera moderna.

1.3.2 Objetivos específicos

- Recolectar y depurar los datos transaccionales de ventas locales y de exportación de la empresa acuícola correspondientes a los períodos comprendidos entre 2021-2024, garantizando la integridad y estructura adecuada para su tratamiento en Google Colab.
- Realizar un análisis exploratorio de los datos mediante técnicas estadísticas y visualizaciones gráficas en Python, con el propósito de identificar distribuciones, correlaciones, tendencias y posibles anomalías en los datos transaccionales.
- Determinar modelos de machine learning que se adapten al sector camaronero, capaz de detectar patrones significativos en los datos comerciales mejorando la auditoría financiera.
- Interpretar y presentar los resultados obtenidos mediante gráficos y reportes generados con Python dentro de Google Colab, con el fin de sustentar hallazgos clave y apoyar procesos de auditoría financiera basados en datos.

1.4. Caracterización del contexto donde se produce/desarrolla el problema o se identifica la oportunidad

A finales de la década de 1960, en Ecuador se inició la actividad de cultivo de camarones, la cual ha experimentado un notable crecimiento y se ha consolidado como un sector clave en la

economía del país. En el presente contexto, la empresa analizada fue fundada en el año 1973 ubicada en la ciudad de Guayaquil, al suroeste de Ecuador en la región costera donde las condiciones naturales son favorables para la camaronicultura. Desde su fundación, ha crecido hasta convertirse en una de las principales exportadoras del sector y ha tenido una participación fundamental en el progreso del sector de la acuicultura de camarones contribuyendo significativamente a la economía nacional con productos de alta demanda en mercados internacionales como Estados Unidos, China, Rusia y la Unión Europea.

La empresa cuenta con operaciones que cubren toda la cadena de valor del camarón, desde el cultivo en piscinas hasta el proceso de empaque y exportación del producto final. Los procesos clave incluyen la reproducción, el engorde en las piscinas, la cosecha y el procesamiento en las plantas empacadoras donde se garantiza la calidad y seguridad del producto.

El enfoque principal de la empresa se centra en las exportaciones, especialmente en los mercados de Norteamérica, Asia y Europa. En estos mercados, el camarón ecuatoriano destaca por su alta calidad y excelente sabor. La empresa ha logrado una producción anual de miles de toneladas métricas de camarón, lo que la sitúa como uno de los principales exportadores del país en términos de resultados. Esta entidad además ha logrado consolidar su posición en el mercado internacional gracias a la demanda constante y a su reputación de calidad, sin embargo, este éxito ha generado la obligación de reforzar sus procesos internos y de cumplir con estándares de transparencia y confiabilidad en la información financiera.

A pesar de que la empresa ha realizado inversiones para mejoar sus instalaciones de procesamiento, el departamento de contabilidad todavía se apoya principalmente en procedimientos manuales y controles convencionales. La plataforma tecnológica empleada en el

ámbito de la gestión contable presenta una falta de capacidades automatizadas para identificar anomalías de manera inmediata y llevar a cabo el análisis predictivo de información financiera. La creciente demanda de prácticas de control más sofisticadas y ágiles plantea un riesgo potencial para la integridad de los registros financieros y la calidad de la auditoría. La aplicación de técnicas de machine learning en el análisis de auditoría se plantea como una estrategia para mejorar la detección de anomalías, aumentar la exactitud en la toma de decisiones y satisfacer los requisitos normativos y comerciales.

Para obtener resolución al problema identificado, los participantes fundamentales abarcan tanto el equipo interno de la empresa como agentes externos, tanto los empleados de contabilidad y auditoría son los más impactados por la carencia de automatización en los procesos de control interno. La participación del personal en la implementación de técnicas de Machine Learning es fundamental para justificar el logro del proyecto.

El respaldo de la dirección ejecutiva y financiera es fundamental para garantizar la obtención de los recursos requeridos destinados a la actualización tecnológica y la formación del personal.

En resumen, la empresa de larga trayectoria en el sector camaronero ecuatoriano se encuentra en un contexto que demanda la modernización de sus procesos contables y de auditoría para mantener su competitividad en un entorno globalizado. La aplicación de técnicas de machine learning en el análisis predictivo de transacciones comerciales ofrece una oportunidad importante para mejorar el control interno, la eficiencia y la transparencia en la gestión financiera, así como para cumplir con las expectativas de los principales actores involucrados.

CAPÍTULO 2: REFERENTES CONCEPTUALES

2.1. Auditoría financiera moderna

La evolución de la auditoría financiera ha pasado de un enfoque basado en comprobación documental hacia una práctica orientada al análisis de riesgos y datos. Actualmente la auditoría no solo se limita a verificar la razonabilidad de estados financieros, sino que también incorpora herramientas tecnológicas para identificar posibles inconsistencias y apoyar la toma de decisiones en la gerencia (Arens, Elder, Beasley, & Hogan, 2018, p. 229). Este enfoque moderno reconoce la importancia del uso de tecnologías emergentes como el análisis predictivo, permitiendo así a los auditores enfocarse en áreas críticas con procedimientos más eficaces.

2.2. Conceptos y principios de Machine Learning aplicados a la auditoría

El machine learning, también conocido como aprendizaje automático, se define como un subcampo de la inteligencia artificial que posibilita a las máquinas adquirir conocimiento a partir de los datos y realizar predicciones o tomar decisiones sin necesidad de ser programadas de forma explícita para cada tarea. Dentro del ámbito de la auditoría financiera, el aprendizaje automático se emplea con el propósito de identificar irregularidades, examinar extensas cantidades de información y pronosticar tendencias comerciales. La habilidad para adquirir conocimientos es especialmente importante en el campo de la auditoría, ya que los auditores necesitan detectar anomalías en información financiera de alta complejidad.

Según (Peñarreta-Angamarca, Torres-Palacios, & Moreno-Narváez, 2024), estudios recientes han evidenciado que la aplicación de técnicas de aprendizaje automático puede incrementar de manera notable la exactitud en la identificación de fraudes y desviaciones contables, sobrepasando las restricciones de los enfoques convencionales. El análisis de patrones en transacciones comerciales posibilita a los auditores la detección de comportamientos atípicos

que podrían sugerir la presencia de actividades fraudulentas. Esto conlleva a una auditoría que se caracteriza por ser más proactiva en lugar de reactiva.

2.3. Inteligencia artificial

Según (Russel & Norvig, 2021), la inteligencia artificial es una rama de la informática que se enfoca en el diseño y desarrollo de sistemas capaces de ejecutar tareas que normalmente requieren inteligencia humana. Estos sistemas operan mediante algoritmos para procesar datos e identificar patrones, generar inferencias y adaptarse a nuevas situaciones sin intervención específica.

En la auditoría, la IA permite automatizar procedimientos analíticos, optimizar la detección de riesgos, reducir sesgos humanos y mejorar el procesamiento de grandes volúmenes de información transaccional

2.4. Análisis exploratorio de datos (EDA)

(Suresh Rao, Vishnu, & Shaik, 2021) afirman que:

> El análisis exploratorio de datos (EDA por sus siglas en inglés) es una etapa fundamental para examinar conjunto de datos, descubrir patrones, detectar anomalías, verificar supuestos y revisar relaciones antes de aplicar modelos predictivos.
>
> Omitir el EDA puede resultar en modelos incorrectos y uso ineficiente de recursos. Dentro de sus beneficios podemos mencionar la limpieza de datos, la selección de variables relevantes y la validación de relaciones estadísticas. (pp. 1-2).

El EDA constituye una fase fundamental para comprender la estructura y comportamiento de las transacciones comerciales en el presente proyecto. Esta exploración proporciona insumos clave para la selección y entrenamiento de modelos de ML, apoyando así

en la auditoría financiera con un enfoque más objetivo, automatizado y orientado a la detección de riesgos.

2.5. Machine Learning aplicado a auditoria financiera

El ML, como subcampo en la inteligencia artificial, se basa en algoritmos que permiten a los sistemas aprender de los datos y hacer predicciones o decisiones sin estar explícitamente programados para cada tarea (Malheiro, Leocádio, & Reis, 2024).

En el ámbito de auditoría el ML puede modernizar los enfoques de revisión al incorporar técnicas supervisadas tales como: regresión y clasificación; y las no supervisadas como: clustering y la detección de outliers. Esto facilita la toma de decisiones basada en datos, además de que estos modelos permiten reducir el riesgo de error humano, también mejora la eficiencia del muestreo, y proporciona alertas tempranas sobre posibles desviaciones significativas.

2.6. Aprendizaje supervisado

Bartz-Beielstein (2024) sostiene que el aprendizaje supervisado es una subcategoría de técnicas de ML en la que el modelo se entrena a partir de un conjunto de datos etiquetados, es decir, con una variable objetivo que ya se ha conocido previamente. El propósito que tiene es de construir una función predictiva capaz de estimar la salida correspondiente a nuevas entradas no observadas. Para esto existen dos tipos principales de modelos: regresión que es cuando la variable de salida es continua; y el modelo de clasificación que es cuando es categórica.

2.7. Regresión lineal

Para James et al,. (2021) la regresión lineal:

> Es una técnica estadística fundamental dentro del aprendizaje supervisado utilizada para modelar la relación entre una variable dependiente continua y una o más variables independientes. Su objetivo es encontrar la mejor combinación lineal de las

variables explicativas que minimice el error cuadrático entre los valores observador y los estimados.

Matemáticamente la regresión lineal se representa de la siguiente manera:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$$

En el contexto del proyecto relacionado con el sector acuícola, esta técnica resulta especialmente útil para analizar y proyectar las ventas locales y exportaciones, ya que permite identificar el impacto de variables clave como los canales de venta, los destinos y el tipo de presentación del producto. Dicho esto, la regresión lineal no solo contribuye a explicar las fluctuaciones históricas en los volúmenes y montos vendidos, sino que también habilita la generación de modelos predictivos que puedan integrarse en procesos de auditoría y planificación comercial basada en datos.

2.8. Interpretación y visualización de resultados

La interpretación y visualización de resultados constituyen las etapas finales del proceso analítico y son esenciales para transformar los hallazgos encontrados en conclusiones comprensibles y útiles para la toma de decisiones. Interpretar adecuadamente implica contextualizar los modelos predictivos utilizados evaluando la relevancia en el proyecto. Por otro lado, la visualización de datos se refiere a la representación gráfica de información cuantitativa mediante herramientas gráficas como gráficos de dispersión, series temporales o dashboards interactivos, estos nos facilitan la comprensión de tendencias, patrones y anomalías (Levy, 2021).

Para el presente proyecto, analizando particularmente las ventas locales y las exportaciones, la interpretación de resultados permite identificar factores que afectan el comportamiento de la demanda, como los destinos o tipo de producto. Complementariamente, la

visualización permite detectar patrones de consumo por los dos mercados mencionados y validar el rendimiento de modelos de predicción.

2.9. Explicación teórica y justificación del modelo utilizado para el desarrollo del proyecto o de la intervención

En machine learning existen dos tipos de modelos aplicables a auditoría financiera tales como los modelos supervisados y los no supervisados ya que son los más útiles para la revisión del presente trabajo de los cuales se utilizará solo un modelo supervisado, sin embargo, se detallará comparaciones entre más modelos que también pueden ser aplicables y que también pueden ser objeto de estudio para auditoría financiera.

Para el presente proyecto existen diversos modelos que pueden aplicarse poder cumplir con la identificación de patrones y anomalías en los datos transaccionales, tales como:

2.9.1. Modelos de Inteligencia Empresarial (BI)

Estos modelos permiten la visualización y análisis descriptivo de grandes volúmenes de datos ya que su procesamiento consiste en extracción, transformación y carga de datos (ETL)(Reyes Sarmiento, 2022).

2.9.2. Modelos estadísticos clásicos

Los modelos de regresión lineal y múltiple son útiles para establecer relaciones entre variables, pero limitados frente a la complejidad y no linealidad de grandes bases de datos (Morán Quispe, 2022).

2.9.3. Modelos de Machine Learning

Estos modelos ofrecen una alternativa más meticulosa al permitir identificar patrones complejos, predecir comportamientos, detectar outliers y en el caso de este trabajo automatizar la toma de decisiones en auditoría

En el contexto del sector acuícola, el uso de datos y registros estadísticos para la predicción de eventos sobre la cantidad de libras que se piensan vender o la cantidad de dólares que han logrado ingresar nos puede dar una idea del comportamiento futuro de estos eventos.

En el estudio realizado por (Cruz Ramírez et al., 2023) donde analizaron a una empresa de alimentos balanceados, en el que emplearon un modelo de redes neuronales Long-Short Term Memory (LSTM por sus siglas en inglés) para encontrar patrones complejos en los datos. El modelo les ayudó a predecir las ventas por cliente y por producto, logrando así mejorar la gestión en la cadena de suministro y las existencias. El proyecto demostró el buen resultado que se obtiene con ML y series de tiempo que conducen a las operaciones de la empresa que sean más eficientes y rentables.

(Galvez Ferrerira et al., 2022) realizaron un estudio profundo en Bucaramanga para predecir mejor los delitos de esa ciudad utilizando modelos espaciales de grafos semanales. El modelo que mas se adaptó a este tipo de investigación es un KNN (K-Nearest Neighbor) de clasificación, que tuvo un 59% de exhaustividad y más de 60% de exactitud.

La era del machine learning representa una transformación y revolución significativa en la forma en que se procesan, analizan e interpretan los datos en el ámbito empresarial y financiero. En el ámbito de la auditoría, el machine learning trabaja como una herramienta estratégica para mejorar la precisión, eficiencia y alcance de los procedimientos de revisión.

Es necesario tener en cuenta los múltiples factores que pueden influir en las ventas de camarón, factores externos como los cambios climáticos, tendencias de mercado, entorno político. Estos factores, le suman una capa de complejidad al desarrollar un pronóstico adecuado.

Para abordar el problema que se ha planteado, en este trabajo se va a optar por utilizar los modelos metodológico CRISP-DM (Cross-Industry Standard Process for Data Mining) y EDA

(Exploratory Data Analysis), que se integran en fases como la de comprensión de datos ya que ayuda a explorar el dataset desde un punto de vista descriprivo y visual; y en la fase de preparación de datos los hallazgos del EDA guían decisiones en la selección de variables, esto se complementa con técnicas de machine learning supervisados y no supervisados con la ayuda de Python dentro de Google Colab.

Con la ayuda de Google Colab, podemos evaluar la calidad de los datos detectando errores, duplicados, etc. Tambien lograremos caracterizar distribución de las variables usando herramientas estadísticas como la media, mediana y varianza; además de las visuales como histogramas o boxplots.

Finalmente, para cumplir con los objetivos del EDA exploraremos relaciones entre variables usando correlaciones y diagramas de dispersión para así formular la hipótesis sobre posibles factores predictivos.

En la ilustración 1, se muestran la secuencia de los pasos a seguir del modelo con iteraciones que se pueden realizar según sea necesario y el orden de ejecución de los pasos se puede invertir.

Usando este modelo permite estructurar el proyecto en seis fases bien definidas y replicables que son complementadas con algoritmos de ML como regresión lineal y clustering que van a permitir detectar patrones de comportamiento en las ventas locales y exportaciones e identificar anomalías y posibles fraudes en las transacciones.

Ilustración 1:

Fases de metodología CRISP-DM

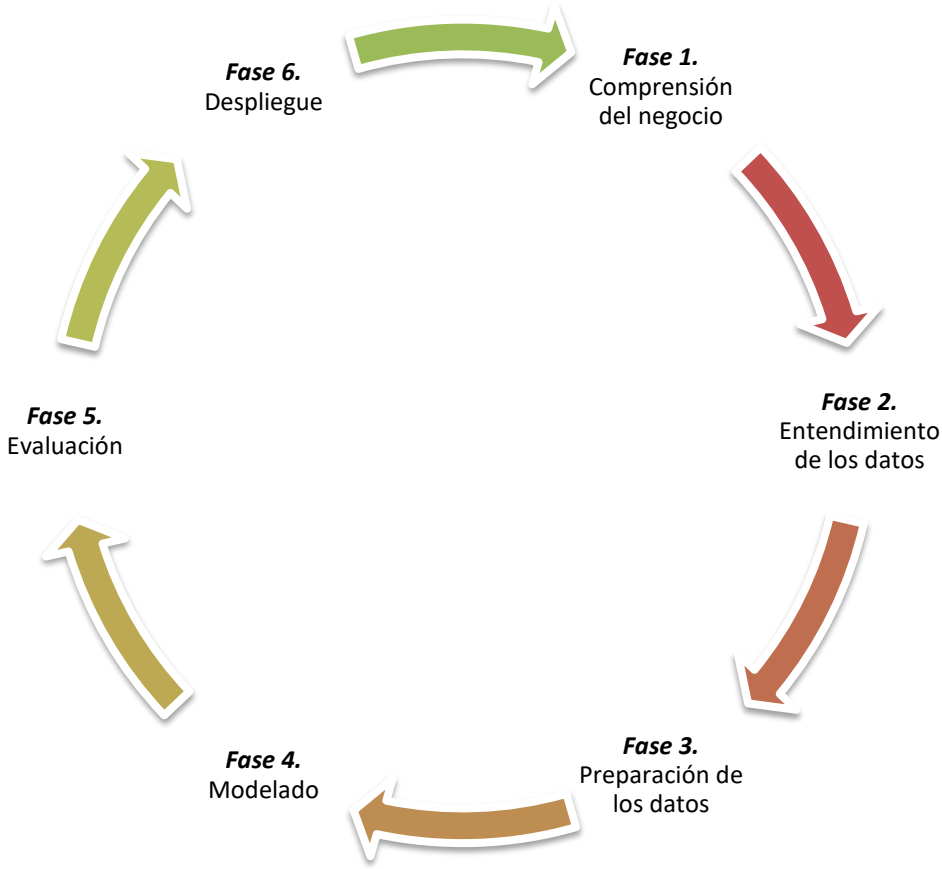


Ilustración 2:

Fases del análisis EDA

CAPÍTULO 3: METODOLOGÍA DE RECOLECCIÓN DE INFORMACIÓN QUE SOPORTA LA PROPUESTA

3.1. Actores/participantes de interés para la propuesta

Los involucrados en el proceso de aporte de información son:

El equipo contable y financiero quienes serán los proveedores de los datos y validación de las transacciones por ventas locales y exportaciones.

El equipo de informática, quienes serán el apoyo para la extracción correcta de la data, y si existe algún inconveniente o error en los datos descargados.

El tutor académico quien será el destinatario de los hallazgos encontrados que causen impacto en los procesos de integración usando machine learning.

3.2. Técnicas de recolección de información/datos

Para el presente proyecto se va a recurrir a una fuente de datos específica para realizar el análisis. Para esto se va a necesitar la base de datos histórica de las transacciones de ventas que van desde el rango de años 2021 – 2024 extraídos del sistema contable de la entidad a consultar.

Para justificar el uso de estos datos resulta estratégico realizar análisis predictivos cuando se dispone de grandes volúmenes de información transaccional almacenada históricamente y registrada de manera sistemática.

Cabe destacar que el objetivo central de esta tesis es aplicar modelos de Machine Learning sobre las ventas locales y exportaciones reales; la fuente más fiable y eficiente es la base de datos generada directamente por el sistema contable de la compañía exportadora del sector acuícola.

Este enfoque es coherente con la naturaleza de los modelos supervisados, los cuales requieren datos históricos con encabezados clave para el entrenamiento, evaluación y validación de modelos predictivos.

Las variables que se van a utilizar son la fecha de emisión, tipo de producto, código del cliente, canal de venta, provincia y país de destino, precio unitario, cantidad y libras vendidas, subtotales, vendedor asignado y tipo de documento.

El formato original de esta base de datos va a ser en formato .xlsx que va a ser exportado del sistema ERP y el volumen estimado de registros por cada archivo anual es mayor a 20,000 líneas que van a ser procesadas en Python dentro de Google Colab.

Como consideración ética el uso de estas fuentes respetará la confidencialidad y anonimato de los datos de la empresa, eliminando información sensible como el RUC del cliente.

Los datos procesados serán utilizados únicamente para fines académicos para el entrenamiento del dataset.

3.3. Plan de recolección y análisis de la información

Tabla 1

Cronograma de recolección y análisis de la información

Actividad	Responsable	Duración	Fecha estimada	Objetivo de la actividad
Solicitud y autorización de uso de data histórica	Tesista / Contabilidad General	5 días	8 de Enero – 12 de Enero 2025	Obtener formalmente los permisos de accesos a base de datos de 2021 a 2024
Exportación del dataset bruto del sistema contable	Área de TI / Tesista	3 días	13 de Enero – 15 de Enero 2025	Extraer la base de datos en formato .xlsx con las variables definidas
Revisión de estructura y limpieza inicial del dataset	Tesista	10 días	16 de Enero – 27 de Enero 2025	Identificar valores faltantes, errores de codificación y duplicados
Validación de variables con usuarios del sistema	Tesista	5 días	28 de Enero – 3 de Febrero 2025	Confirmar el significado operativo de cada campo clave (producto, canal, cliente, etc.)
Análisis exploratorio de datos (EDA)	Tesista	10 días	4 de Febrero – 14 de Febrero 2025	Evaluar estadísticamente las variables y definir comportamientos relevantes
Selección del modelo y definición de etiquetas	Tesista	7 días	15 de Febrero - 21 de Febrero 2025	Determinar si se aplicará regresión o clasificación, y qué variable será el objetivo
Entrenamiento de modelos en Python	Tesista	14 días	22 de Febrero - 7 de Marzo 2025	Entrenar, ajustar y evaluar modelos predictivos
Validación de resultados	Tesista / Tutor de Tesis	5 días	8 de Marzo – 12 de Marzo 2025	Validar si los resultados son coherentes
Interpretación y elaboración de reportes	Tesista	10 días	13 de Marzo – 23 de Marzo 2025	Documentar hallazgos, visualizaciones y patrones identificados.
Redacción de capítulo de resultados y discusión	Tesista	15 días	24 de Marzo – 8 de Abril 2025	Integrar resultados al cuerpo del trabajo con redacción académica.

CAPÍTULO 4: RESULTADOS

4.1. Fase 1. Entendimiento del negocio

En la primera fase de la metodología CRISP-DM que consiste en la comprensión del negocio, la empresa de la que es objeto este trabajo se dedica al cultivo, distribución y venta de camarón y tilapia, siendo su fuerte el camarón en el mercado nacional e internacional.

Para esto se recolectó archivos históricos en formato Excel, que contenían información sobre las operaciones de venta. Los archivos están separados en dos conjuntos principales que son ventas locales y exportaciones. A través de técnicas de procesamiento de datos con *pandas*, para esto se realizó a recolectar, almacenar y consolidar la data histórica en Google Drive, mediante la utilización de Google Colab como entorno de desarrollo.

En la figura 2 se identifican dos fuentes de información que son las exportaciones, en esta incluyen atributos como producto, kilos netos, libras netas, valor FOB, flete, país, cliente, ciudad destino y tipo de transporte.

En la segunda figura 1 son las ventas locales que están compuestas por información detallada como tipo de producto, canales de venta, libras, subtotal, total, ciudad, provincia y vendedor.

Desde un punto de vista crítico, el diseño de la base de datos es adecuada para análisis descriptivos, predictivos y de detección de anomalías. Existen una diversidad de variables que refuerza la aplicabilidad de técnicas de machine learning supervisadas y no supervisadas.

Finalmente, se identificó algunos aspectos clave que van a ser abordados en la preparación de estos datos que van a ser considerados en la preparación de los datos:

Homogeneizar los formatos de las fechas y tipos de variables categóricas.

La presencia de valores nulos o inconsistencias menores que deben ser tomadas en cuenta para evitar sesgos.

La oportunidad de crear variables derivadas como las ventas netas o la clasificación por región y cliente.

Figura 1:

Dataset ventas locales

	TIPO DE PRODUCTO	LINEA DE PRODUCTO	COD ITEM	DESCRIPCION_ITEM	DESCRIPCION_CORTA	SERIE FC	FACTURA	FECHA EMISION	AÑO	CANAL DE VENTA	...
0	CAMARON	CAMARON VENTA LOCAL	9486	VENTAS ECUADOR_FDAS X 5 LBS._26-30_IQF P & D ...	IQF P & D CRUDO LLANO 1ERA 26-30 5 LBS FDAS	1014	1014000032016	2021-01-14	2021	HORECA	...
1	CAMARON	CAMARON VENTA LOCAL	9486	VENTAS ECUADOR_FDAS X 5 LBS._26-30_IQF P & D ...	IQF P & D CRUDO LLANO 1ERA 26-30 5 LBS FDAS	1014	1014000034934	2021-04-15	2021	HORECA	...
2	CAMARON	CAMARON VENTA LOCAL	9486	VENTAS ECUADOR_FDAS X 5 LBS._26-30_IQF P & D ...	IQF P & D CRUDO LLANO 1ERA 26-30 5 LBS FDAS	1014	1014000038232	2021-07-16	2021	HORECA	...
3	CAMARON	CAMARON VENTA LOCAL	9486	VENTAS ECUADOR_FDAS X 5 LBS._26-30_IQF P & D ...	IQF P & D CRUDO LLANO 1ERA 26-30 5 LBS FDAS	1014	1014000042745	2021-11-15	2021	HORECA	...
4	CAMARON	CAMARON VENTA LOCAL	9788	VENTAS ECUADOR (SIN TRATAMIENTO)_FDAS X 5 LBS...	IQF P & D CRUDO LLANO 1ERA 31-35 5 LBS FDAS	1006	1006000030164	2021-10-14	2021	MAYORISTA	...

5 rows × 23 columns

Figura 2:

Dataset exportaciones

	Desc. Producto	TIPO DE PRODUCTO	Talla	No. Embarque	Serie Factura	Num. Factura	Fecha Factura	Fecha Embarque	Libras Netas	Kilos Netos	Valor FOB	Flete	CFR	Cod. Cliente	Cliente
0	CAJAS VACIAS	MATERIALES / OTROS	CAJAS/TROPICAL	9924	1008	2225	2022-01-06	2022-01-06 15:37:37	140.0	63.503583	48.270040	91.73	140.000040	90	TROPICAL AQUACULTURE PRODUCTS, INC.
1	CAJAS VACIAS	MATERIALES / OTROS	CAJAS/TROPICAL	9940	1008	2241	2022-02-10	2022-02-10 09:24:46	140.0	63.503583	43.449980	96.55	139.999980	90	TROPICAL AQUACULTURE PRODUCTS, INC.
2	CAJAS VACIAS	MATERIALES / OTROS	CAJAS	9958	1008	2259	2022-03-14	2022-03-14 14:15:03	114.0	51.710061	29.659950	84.34	113.999950	90	TROPICAL AQUACULTURE PRODUCTS, INC.
3	CAMARON CONGELADO	CAMARON	41-50	52762	1007	39895	2022-12-23	2022-12-29 06:24:00	5250.0	2381.400000	10163.474898	552.83	10716.304898	1334	FUZHOU FASHIONL TECHNOLOGY CO., LTD
4	CAMARON CONGELADO	CAMARON	26-30	52768	1007	39901	2022-12-21	2022-12-29 23:00:00	15980.0	7248.528000	47004.633176	2533.37	49538.003176	151	PACIFIC BREEZE SEAFOOD INC

4.2. Fase 2. Entendimiento de los datos

La base de datos a utilizar consiste en reportes descargados de las ventas registradas en los años 2021 al 2024 provenientes del sistema de la entidad anónima a estudiar; la comprensión de los datos comprende una fase fundamental para garantizar la calidad del modelo predictivo, en especial en el contexto de la auditoría.

4.2.1. Dimensiones del dataset

Figura 3:

Dimesión del dataset ventas locales

Dimensiones del dataset: 882,567 filas × 23 columnas

Estructura del dataset y porcentaje de valores nulos

	Columna	Tipo de dato	Nulos	% Nulos
0	TIPO DE PRODUCTO	object	0	0.00%
1	LINEA DE PRODUCTO	object	0	0.00%
2	COD ITEM	int64	0	0.00%
3	DESCRIPCION_ITEM	object	0	0.00%
4	DESCRIPCION_CORTA	object	0	0.00%
5	SERIE FC	int64	0	0.00%
6	FACTURA	int64	0	0.00%
7	FECHA EMISION	datetime64[ns]	0	0.00%
8	AÑO	int64	0	0.00%
9	CANAL DE VENTA	object	0	0.00%
10	ID CLIENTE	int64	0	0.00%
11	CLIENTE	object	0	0.00%
12	PRECIO	float64	0	0.00%
13	CANTIDAD	float64	0	0.00%
14	LIBRAS	float64	0	0.00%
15	SUBTOTAL	float64	0	0.00%
16	IVA	float64	0	0.00%
17	TOTAL	float64	0	0.00%
18	ID VENDEDOR	int64	0	0.00%
19	VENDEDOR	object	0	0.00%
20	TIPO DOC	object	0	0.00%
21	CIUDADES	object	0	0.00%
22	PROVINCIA	object	0	0.00%

Figura 4:

Dimensión del dataset exportaciones

Resumen del DataFrame:

Estructura del dataset y porcentaje de valores nulos

	Tipo de Dato	Valores Nulos	No Nulos
Desc. Producto	object	0	0.000000
TIPO DE PRODUCTO	object	0	0.000000
Talla	object	12051	20.659329
No. Embarque	int64	0	0.000000
Serie Factura	int64	0	0.000000
Num. Factura	int64	0	0.000000
Fecha Factura	datetime64[ns]	0	0.000000
Fecha Embarque	datetime64[ns]	0	0.000000
Libras Netas	float64	0	0.000000
Kilos Netos	float64	0	0.000000
Valor FOB	float64	0	0.000000
Flete	float64	0	0.000000
CFR	float64	0	0.000000
Cod. Cliente	int64	0	0.000000
Cliente	object	0	0.000000
Transporte	object	17	0.029144
Presentacion	object	19	0.032572
MarcaP	object	0	0.000000
pais_esp	object	0	0.000000
pais_eng	object	0	0.000000
ciudad	object	0	0.000000

Dimensiones del dataset de exportaciones: 58,332 filas × 21 columnas

En esta fase de comprensión de datos, se analizó las estructuras de los datasets correspondientes a ventas locales y exportaciones.

El conjunto de ventas locales contiene 882.567 filas con registros que están distribuidos en 23 filas, lo que evidencia una adecuada integridad estructural, especialmente en variables críticas como *PRECIO, CANTIDAD, SUBTOTAL y TOTAL*, fundamentales para análisis.

Por otro lado, el dataset de exportaciones contiene 58.332 registros y 21 filas, donde se puede observar una calidad de datos generalmente alta, aunque existen inconsistencias en la variable "*Talla*", la cual presenta 12.051 casillas nulas, equivalente al 20,65% de los registros.

Estas ausencias podrían constituir un riesgo potencial en la trazabilidad del producto exportado y limita los análisis para la evaluación de ingresos por tipo de presentación.

4.2.2. Valores del dataset

Figura 5:

Variables exportaciones

Valores únicos por columna:

	Columna	Valores Únicos
0	Desc. Producto	11
1	TIPO DE PRODUCTO	5
2	Talla	190
3	No. Embarque	40803
4	Serie Factura	4
5	Num. Factura	40819
6	Fecha Factura	1179
7	Fecha Embarque	13662
8	Libras Netas	6526
9	Kilos Netos	6724
10	Valor FOB	32327
11	Flete	14881
12	CFR	31067
13	Cod. Cliente	488
14	Cliente	509
15	Transporte	4
16	Presentacion	215
17	MarcaP	274
18	pais_esp	45
19	pais_eng	40
20	ciudad	355

Figura 6:

Variables categóricas

```
# Verificar valores únicos en variables categóricas
print("Productos:", df["Desc. Producto"].unique())
print("Países:", df["pais_esp"].unique())
print("Tipo de Producto:", df["TIPO DE PRODUCTO"].unique())
print("Clientes:", df["Cliente"].nunique())
```

```
Productos: ['CAJAS VACIAS' 'CAMARON CONGELADO' 'CAMARON FRESCO' 'ENTERO FRESCO'
 'FILETE CONGELADA' 'FILETE FRESCO' 'GELPACK' 'HARINA DE CAMARON'
 'HAMBURGUESA DE CAMARON' 'SIN DESCRIPCION' 'OTROS']
Países: ['USA' 'TAILANDIA' 'ITALIA' 'BELGICA' 'JAPON' 'FRANCIA' 'RUSIA' 'VIETNAM'
 'UCRANIA' 'CHILE' 'ESPAÑA' 'MARRUECOS' 'SOUTH AFRICA' 'HOLANDA'
 'COLOMBIA' 'PORTUGAL' 'EMIRATOS ARABES UNIDOS' 'POLONIA' 'SINGAPORE'
 'GRECIA' 'PERU' 'CANADA' 'CHINA' 'COREA DEL SUR' 'REINO UNIDO' 'MALASIA'
 'Holanda' 'GUATEMALA' 'COSTA DE MARFIL' 'IRELAND' 'SUDAFRICA' 'LITUANIA'
 'RUMANIA' 'COSTA RICA' 'NORUEGA' 'PUERTO RICO' 'South Africa' 'ALEMANIA'
 'MEXICO' 'SRI LANKA' 'ARGENTINA' 'PANAMA' 'SRI LANKA (CEILAN)' 'TAIWAN'
 'JORDANIA']
Tipo de Producto: ['MATERIALES / OTROS' 'CAMARON' 'PESCADO' 'SIN DESCRIPCION' 'FORMULACION']
Clientes: 509
```

Como parte del proceso de comprensión del dataset, en las figuras 5 y 6 se analizaron las variables categóricas clave asociadas a las exportaciones: productos, países de destino, tipos de producto y clientes. El objetivo de este análisis fue identificar la diversidad, consistencia y posibles problemáticas estructurales en los datos, elementos esenciales para una posterior limpieza, transformación y análisis predictivos confiable.

En los 45 países encontrados existe una variante léxica con Sri Lanka, lo que podría conducir a duplicaciones o errores en los análisis por país. Para esto se efectuará una homogeneización de las nomenclaturas en la fase de preparación de datos.

Figura 7:

Valores únicos venta local

Valores únicos por columna:

	Columna	Valores Únicos
0	TIPO DE PRODUCTO	6
1	LINEA DE PRODUCTO	63
2	COD ITEM	2128
3	DESCRIPCION_ITEM	2274
4	DESCRIPCION_CORTA	2124
5	SERIE FC	22
6	FACTURA	417160
7	FECHA EMISION	1444
8	AÑO	4
9	CANAL DE VENTA	11
10	ID CLIENTE	41201
11	CLIENTE	40977
12	PRECIO	3875
13	CANTIDAD	9590
14	LIBRAS	10740
15	SUBTOTAL	33592
16	IVA	7710
17	TOTAL	36345
18	ID VENDEDOR	25
19	VENDEDOR	27
20	TIPO DOC	7
21	CIUDADES	8
22	PROVINCIA	8

Figura 8:

Verificación duplicados ventas locales

Valores faltantes por columna:

	Columna	Valores Faltantes
0	TIPO DE PRODUCTO	0
1	LINEA DE PRODUCTO	0
2	COD ITEM	0
3	DESCRIPCION_ITEM	0
4	DESCRIPCION_CORTA	0
5	SERIE FC	0
6	FACTURA	0
7	FECHA EMISION	0
8	AÑO	0
9	CANAL DE VENTA	0
10	ID CLIENTE	0
11	CLIENTE	0
12	PRECIO	0
13	CANTIDAD	0
14	LIBRAS	0
15	SUBTOTAL	0
16	IVA	0
17	TOTAL	0
18	ID VENDEDOR	0
19	VENDEDOR	0
20	TIPO DOC	0
21	CIUDADES	0
22	PROVINCIA	0

Columnas disponibles:

Nombre de Columna

No hay columnas duplicadas exactamente.

No hay columnas con nombres inconsistentes detectadas.

Por el lado de venta local se comparte el mismo objetivo que es la de garantizar la calidad de los datos empleados para el modelo predictivo y se procedió a analizar de manera exploratoria con enfoque en dos aspectos que son los valores únicos y la ausencia de valores nulos por variable.

En ventas locales no se detectaron novedades con columnas duplicadas o inconsistentes lo que facilita la automatización en el procesamiento del dataset.

4.3. Fase 3: Preparación de los datos

En esta fase se ejecutó el proceso de depuración, transformación y estructuración de los datos con el propósito de garantizar su calidad, coherencia y utilidad para el modelado posterior. El conjunto de datos compuesto por registros de ventas locales y exportaciones entre los años 2021 y 2024 fue sometido a procedimientos de limpieza, incluyendo conversión de fechas a formato estándar, detección y eliminación de valores nulos.

Para asegurar la trazabilidad de los datos, se documentaron todas las transacciones realizadas en el entorno de Google Colab con Python, preservando la integridad de la fuente original.

Finalmente, se consolidó un único DataFrame por segmento como lo es el local y el de exterior, preparado para ser utilizado en el desarrollo de modelos de machine learning en las siguientes fases.

Figura 9:

Verificación de limpieza de datos exportaciones

```
from IPython.display import display

# Asegurarse que las columnas tengan nombres limpios
df.columns = df.columns.str.strip()

# Forzar conversión a string para procesar caracteres y luego a float
for col in ['Valor FOB', 'Flete', 'CFR','Libras Netas', 'Kilos Netos']:
    df[col] = df[col].astype(str) \
                     .str.replace('[^0-9.,-]', '', regex=True) \
                     .str.replace(',', '.') \
                     .astype(float)

# Convertir FECHA EMISION a formato datetime
df['Fecha Factura'] = pd.to_datetime(df['Fecha Factura'], errors='coerce', dayfirst=True)

# Verificación rápida en forma de tabla
print("\nVerificación de conversión de columnas:")
display(df[['Fecha Factura', 'Valor FOB', 'Flete', 'CFR', 'Libras Netas', 'Kilos Netos']].head())
```

```
Verificación de conversión de columnas:
```

	Fecha Factura	Valor FOB	Flete	CFR	Libras Netas	Kilos Netos
0	2021-03-27	22.709995	38.29	60.999995	61.0	27.669418
1	2021-04-08	48.419944	87.58	135.999944	136.0	61.689195
2	2021-04-22	47.889952	88.10	135.989952	136.0	61.689195
3	2021-05-17	91.900120	156.10	248.000120	248.0	112.492062
4	2021-06-14	78.890112	209.11	288.000112	288.0	130.635943

Figura 10:

Verificación de limpieza de datos ventas locales

```
from IPython.display import display

# Asegurarse que las columnas tengan nombres limpios
df.columns = df.columns.str.strip()

# Forzar conversión a string para procesar caracteres y luego a float
for col in ['PRECIO', 'CANTIDAD', 'LIBRAS', 'SUBTOTAL', 'TOTAL']:
    df[col] = df[col].astype(str) \
                     .str.replace('[^0-9.,-]', '', regex=True) \
                     .str.replace(',', '.') \
                     .astype(float)

# Convertir FECHA EMISION a formato datetime
df['FECHA EMISION'] = pd.to_datetime(df['FECHA EMISION'], errors='coerce', dayfirst=True)

# Verificación rápida en forma de tabla
print("\nVerificación de conversión de columnas:")
display(df[['FECHA EMISION', 'CANTIDAD', 'LIBRAS', 'PRECIO', 'SUBTOTAL', 'TOTAL']].head())
```

```
Verificación de conversión de columnas:
```

Columna	FECHA EMISION	CANTIDAD	LIBRAS	PRECIO	SUBTOTAL	TOTAL
0	2021-01-14	400.0	400.0	5.30	2120.0	2120.0
1	2021-04-15	400.0	400.0	5.30	2120.0	2120.0
2	2021-07-16	400.0	400.0	5.30	2120.0	2120.0
3	2021-11-15	400.0	400.0	5.30	2120.0	2120.0
4	2021-10-14	20.0	20.0	2.88	57.6	57.6

En este paso se procesaron las columnas de exportaciones *Valor FOB, Flete, Libras netas, Kilos netos* y *CFR*, las cuales inicialmente no contenían caracteres numéricos. Se aplicó una estrategia de limpieza basada en expresiones regulares para extraer únicamente los valores válidos, seguidos de la conversión a tipo *float,* esto garantizó una precisión para validar la integridad de las transacciones.

Adicionalmente, la columna *Fecha Factura* fue transformada a tipo *datetime*, habilitando el análisis temporal mediante la creación de variables como mes, año o día de la semana. Esta transformación es esencial para analizar las estacionalidades y transacciones recurrentes.

En mercado local se aplicó la misma técnica para las columnas *PRECIO, SUBTOTAL, LIBRAS, CANTIDAD*, que inicialmente estaban en formato texto. Posteriormente la columna *FECHA EMISION* fue convertida al formato *datetime.*

En conclusión, este proceso permitió consolidar los datos de manera estructurada, coherente y alineado a estándares analíticos, tanto para exportaciones como para mercado local.

Culminada esta fase, los datos están listos para su explotación en la siguiente fase, permitiendo aplicar técnicas de machine learning, análisis exploratorio y detección de patrones atípicos.

4.4. Fase 4: Modelado

4.4.1. Modelos no supervisados

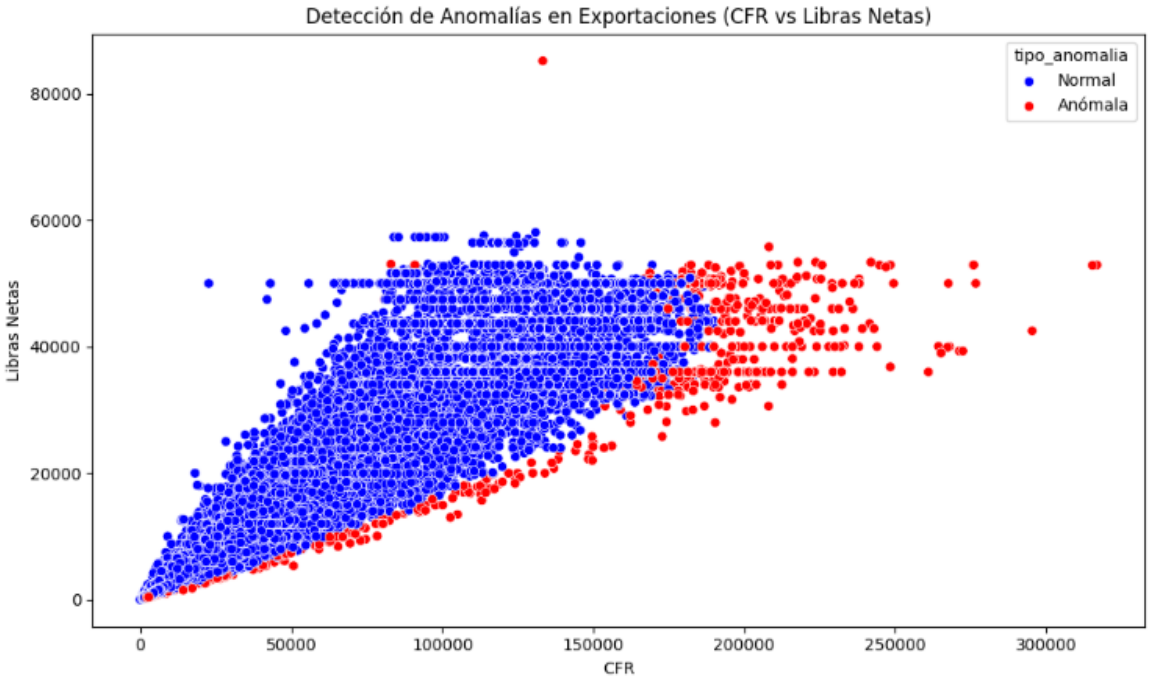
En el presente trabajo se entrenó ciertos modelos de machine learning entre supervisados y no supervisados con el fin de detectar valores atípicos en el conjunto de datos preprocesados.

4.4.1.1. Isolation Forest

Para este análisis se aplicó un modelo no supervisado para la detección de anomalías utilizando el algoritmo Isolation Forest, ya que esta técnica es muy utilizada en auditoría forense y control financiero, esta permite identificar observaciones atípicas que difieren sustancialmente de los patrones normales en los datos.

Figura 11:

Bosque de aislamiento exportaciones



En la figura 11 Se representa un gráfico de dispersión con las variables CFR y Libras Netas en las exportaciones. Cada punto representa una transacción de exportación, y se utiliza una señal de color para distinguir entre transacciones normales y anómalas por el modelo:

Las esferas de color azul representan transacciones están que clasificadas como comportamientos regulares por el modelo

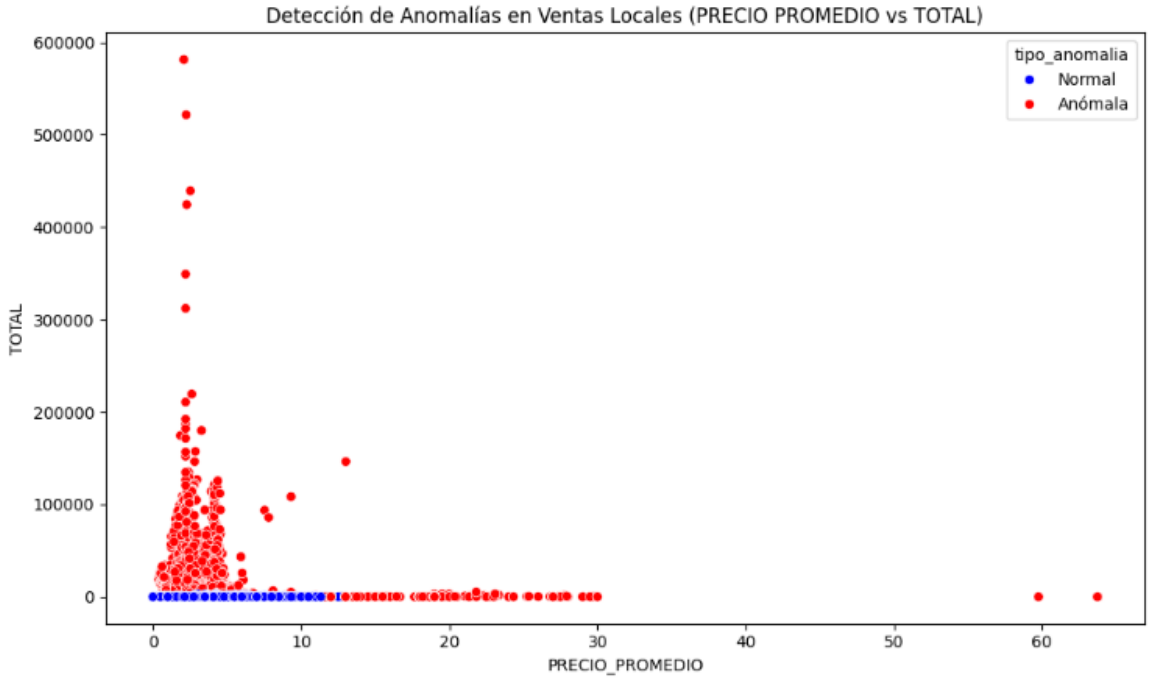
Las esferas de color rojo representan transacciones que fueron clasificadas como atípicas con base en su divergencia respecto al patón general.

El gráfico muestra mayoría de transacciones que siguen una relación aproximadamente lineal entre las libras exportadas y el CFR, lo cual es normal. Sin embargo, existen un grupo de puntos rojos que se dispersa fuera del patrón denso principal, indicando transacciones con proporciones inusuales entre peso y valor.

Estos valores atípicos pueden estar asociadas a errores de registro, facturación atípica, manipulación o irregularidad en datos comerciales.

Figura 12:

Bosque de aislamiento ventas locales



Para el mercado local se aplicó el mismo algoritmo, sin embargo, dado que cada fila del dataset de venta local representa un ítem, los valores numéricos *PRECIO, CANTIDAD, LIBRAS,*

TOTAL, no reflejaban una transacción completa, sino solo una parte de ella, por lo que se optó por agrupar las filas por factura y sacamos un precio promedio y total de la factura. El gráfico obtenido representa en el eje ‘X’ el precio promedio por libra y en el eje ‘Y’ el valor total de la factura, donde cada punto corresponde a una transacción agregada.

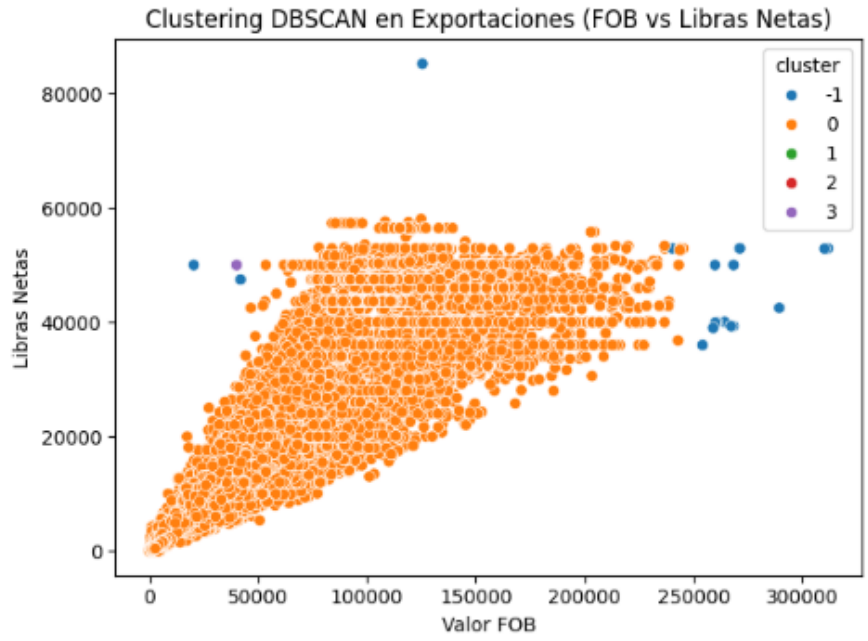
El modelo identificó como anómala aquellas facturas cuyos valores se alejaban significativamente del patrón central observado en la nube de puntos normales. Se destaca que muchas de las anomalías se encuentran en extremos del eje ‘X’, es decir, en facturas con precios promedio atípicamente bajos o altos, y en el eje ‘Y’, con montos considerablemente elevados o reducidos, lo que podría ser indicativo de posibles duplicaciones de ítems, o una mala asignación de cantidades. Estos análisis son relevantes para procesos de auditoría financiera moderna, ya que además de identificar irregularidades cuantitativas, también ofrece un mecanismo automatizado para mejorar la trazabilidad y el control de las ventas locales y exportaciones.

4.4.1.2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

Para observar agrupaciones naturales en las transacciones de exportación, se implementó el algoritmo de clustering DBSCAN. Este modelo no supervisado es particularmente adecuado para identificar agrupaciones de densidad variable y detectar valores atípicos o ruido en conjuntos de datos complejos.

Figura 13:

Algoritmo DBSCAN exportaciones




```
outliers = df_export_clean[df_export_clean['cluster'] == -1]
print(f"Número de puntos etiquetados como ruido (anómalos): {len(outliers)}")
```

```
Número de puntos etiquetados como ruido (anómalos): 26
```

Se utilizó como referencia las variables Valor FOB y Libras Netas, representativas del volumen y valor de las exportaciones. En el gráfico generado, los distintos colores indican los grupos o clústeres detectados, mientras que los puntos identificados como *ruidos* representados con *-1* representan posibles transacciones atípicas que no se ajustan a ninguna agrupación densa.

El modelo identificó cuatro agrupaciones principales (clusters 0 al 3) y 26 observaciones fueron etiquetadas como ruido. Estos puntos anómalos representan casos que podrían corresponder a registros erróneos, transacciones fuera de los rangos esperados, lo que justifica un análisis más profundo.

Desde una perspectiva de auditoria financiera moderna, la detección de estos puntos es esencial para alertar sobre posibles errores de facturación, identificar exportaciones

potencialmente irregulares ya sea por sobrevaloración, subvaloración o manipulación de cantidades.

Figura 14:

Algoritmo DBSCAN Ventas locales 2021

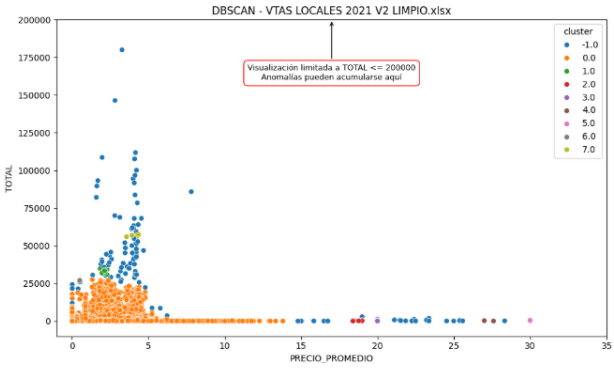


Figura 15:

Algoritmo DBSCAN Ventas Locales 2023

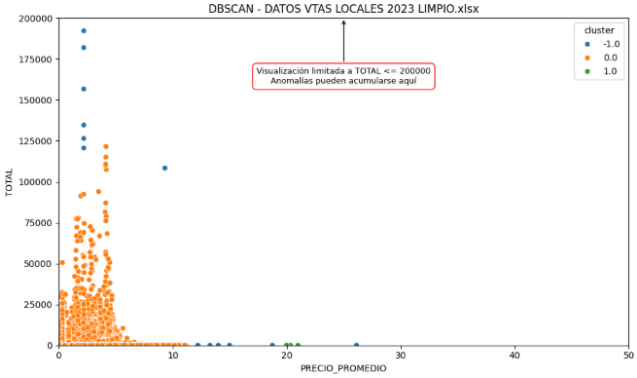


Figura 17:

Algoritmo DBSCAN Ventas locales 2022

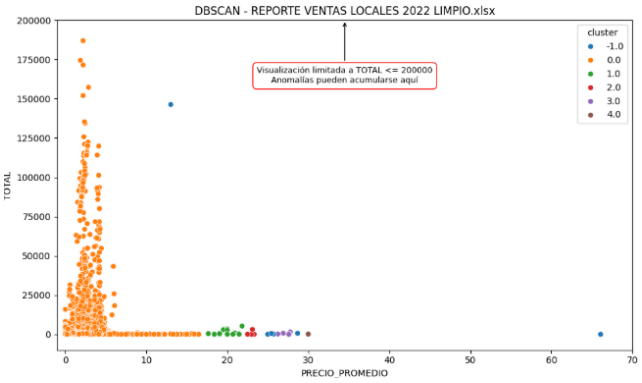
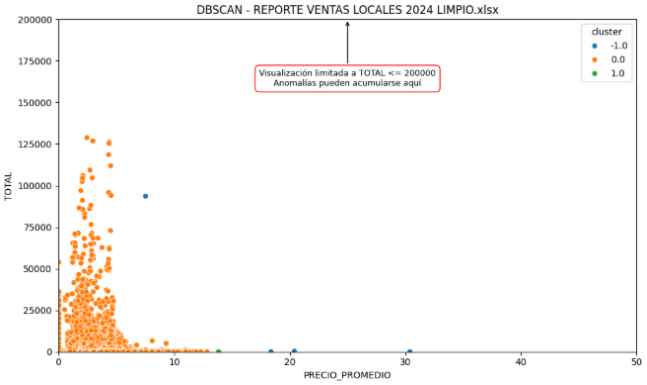


Figura 16:

Algoritmo DBSCAN Ventas locales 2024



Para venta local se realizó el análisis de manera individual debido al volumen de datos que contiene la base. En 2021 hubo una alta diversidad de transacciones con presencia de valores atípicos (-1 color azul) que reflejan posibles errores o ventas excepcionalmente grandes. En el año 2022 presenta una tendencia más clara en las ventas, con menos variabilidad en comparación

con 2021. En el 2022 las anomalías son más aisladas y fáciles de identificar, lo que sugiere mayor control o estandarización en las operaciones. En 2023 existe alta concentración en un solo clúster (0 color naranja) y pocos puntos etiquetados como otro grupo o anomalía. Este año muestra una homogeneidad mas marcada, señal de consolidación del comportamiento comercial. Las anomalías son menos numerosas, lo que podría reflejar procesos internos más depurados. En el año 2024 se mantiene una concentración fuerte en valores bajos de precio y total. Se identifican anomalías visibles, lo que puede indicar estabilidad en los datos. En 2024 parece haber alcanzado una mejoría operativa. Las ventas son más predecibles y con menos desviaciones.

4.4.2. Modelos supervisados

4.4.2.1. Regresión logística (mercado exportaciones)

Se usó la regresión logística como modelo predictivo inicial para identificar patrones en las transacciones comerciales de la empresa acuícola, considerando tanto las exportaciones como las ventas locales.

En el contexto de la auditoría financiera moderna la regresión logística permitió interpretar la influencia de factores como el tipo de transporte, el país de destino y el tipo de producto sobre la probabilidad de que una venta sea significativa para mejorar las revisiones contables.

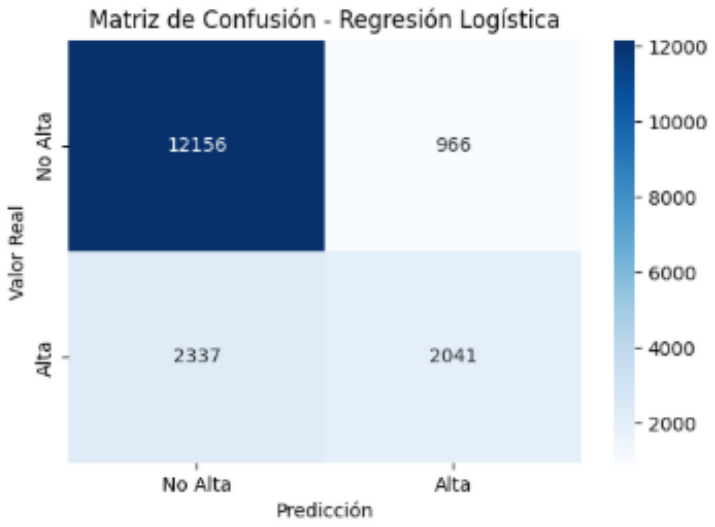
Además, permitió priorizar los riesgos identificando sistemáticamente las transacciones que, por su valor o características logísticas requieren procedimientos analíticos adicionales, alineándose con los enfoques de auditoría basada en riesgos.

Al generar probabilidades para cada registro, facilita la creación de reportes dinámicos que combinan criterios estadísticos con indicadores contables, que contribuyen a la automatización y eficiencia de la revisión analítica.

4.4.2.1.1 Matriz de confusión

Figura 18:

Matriz de confusión-Regresión Logística Exportaciones



La matriz de confusión permitió visualizar la capacidad del modelo para diferenciar entre las ventas altas que representan las clases positivas y las ventas no altas representadas en negativas.

En ellas evidencian:

Los verdaderos positivos que son transacciones correctamente clasificadas como ventas altas.

Los verdaderos negativos son transacciones correctamente identificadas como no altas.

Los falsos positivos son operaciones clasificadas como ventas altas sin serlo, lo que representa un riesgo de sobre detección.

Los falsos negativos son ventas altas que el modelo no logró identificar, representando un riesgo de omisión en la auditoría.

La matriz evidenció un buen equilibrio entre verdaderos positivos y verdaderos negativos, mostrando que el modelo logra identificar correctamente un porcentaje importante de las ventas relevantes, con un número moderado de errores de clasificación.

4.4.2.1.2. Curva ROC (Receiver Operating Characteristic)

Figura 19:

Curva ROC-Regresión Logística Exportaciones

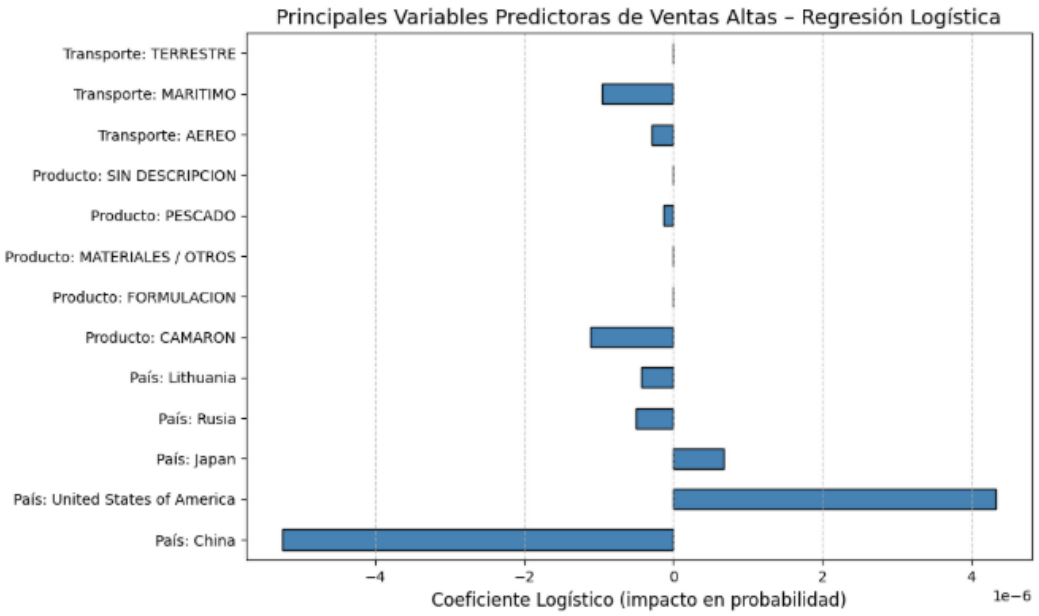
La curva ROC (Receiver Operating Characteristic) graficó la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos.

El área bajo la curva (AUC) obtenida fue de 0.86, lo que indica que el modelo tiene una capacidad de discriminación aceptable a buena para diferenciar entre ventas altas y no altas. Este indicador es especialmente relevante para auditoría financiera basada en riesgos, ya que permite priorizar las operaciones con mayor probabilidad de ser significativas, optimizando los recursos de revisión analítica.

4.4.2.1.3. Coeficiente logístico

Figura 20:

Gráfico de Importancia de Variables en Regresión Logística Exportaciones



En la figura 20 muestra la influencia de los principales factores logísticos y comerciales sobre la probabilidad de que una transacción de exportación sea clasificada como venta alta.

El eje horizontal representa los coeficientes logísticos estimados, que reflejan la dirección e intensidad del impacto de cada variable sobre la probabilidad de ventas altas:

Los coeficientes positivos indican la probabilidad de que la transacción sea clasificada como venta alta.

Los coeficientes negativos indican reducción de dicha probabilidad; y los valores cercanos a cero indican un impacto marginal.

Entre los países destino destacan:

Estados Unidos que presenta un coeficiente positivo significativo, indica que las exportaciones hacia este mercado tienen una mayor probabilidad de convertirse en operaciones de alto valor

China, por su lado, muestra un coeficiente negativo, lo que sugiere que las exportaciones hacia este país, aunque son de alto volumen, tienden a concentrarse en operaciones de menor monto unitario.

En cuanto al tipo de producto, el camarón, aunque es el principal producto de la empresa, muestra un coeficiente ligeramente negativo, indicando que la mayoría de estas transacciones se concentran en montos intermedios. Esto se deduce que las ventas mas significativas suelen estar asociadas a productos con valor agregado o formulaciones especiales, que elevan el monto total por operación.

Finalmente, el análisis del modo de transporte evidencia que el transporte aéreo tiene un efecto positivo, asociado a operaciones de mayor valor y urgencia logística, mientras que el transporte marítimo y terrestre presentan coeficientes negativos, ya que existen envíos de gran volumen, pero menor valor unitario.

Con este análisis permite concluir que la regresión logística aplicada a las exportaciones de la entidad camaronera estudiada no solo identifica las transacciones de mayor valor económico, sino que también revela patrones comerciales relevantes para la auditoría financiera moderna, y facilita la prioridad de revisión y enfoque de procedimientos analíticos en mercados, productos y operaciones logísticas de mayor impacto.

4.4.2.1.4. Classification report

Figura 21:

Classification report exportaciones

	precision	recall	f1-score	support
0	0.839	0.926	0.880	13122.000
1	0.679	0.466	0.553	4378.000
accuracy	0.811	0.811	0.811	0.811
macro avg	0.759	0.696	0.717	17500.000
weighted avg	0.799	0.811	0.798	17500.000

El classification report presentado en la figura 21 muestra el desempeño del modelo de egresión logística en la clasificación de transacciones de exportaciones como ventas altas (1) y ventas no altas (0).

Clase 0 - Ventas no altas.

Tuvo una precisión del 83.9% de todas las transacciones clasificadas como no altas.

En el recall el modelo identificó correctamente el 92.6% de todas las transacciones realmente no altas.

El F1-score de 0.88 muestra un alto equilibrio entre precisión y recall , confirmando que el modelo es confiable para descartar operaciones de bajo valor económico.

Clase 1 – Ventas altas

Tiene una precisión de 0.679, esto quiere decir que de todas las transacciones clasificadas como ventas altas, el 67.9% fueron correctas, evidenciando un riesgo moderado de falsos positivos.

En el recall solo el 46.6% de las ventas altas reales fueron detectadas, lo que indica que existe un número relevante de flasos negativos que podrían implicar omisión de transacciones significativas.

EL F1-score fue de 0.553 reflejando un desempeño moderado en la detección de ventas altas, impactando principalmente por la baja sensibilidad.

4.4.2.1.5. Métricas globales

Accuracy o exactitud global del modelo clasifica correctamente 81.1% de las transacciones.

Promedio macro (Macro avg) tiene una precisión de 0.759 y un recall de 0.696 lo que representa el desempeño promedio equilibrado de ambas clases sin ponderar por cantidad de registros.

Weighted promedio (weighted avg) indica que modelo tiene una precisión igual a 0.799 y un recall igual a 0.811 manteniendo un rendimiento global sólido, influenciado por la mayor cantidad de transacciones no altas.

Estos resultados reflejan que el modelo es altamente confiable para identificar ventas no significativas, reduciendo la carga de revisión en operaciones de bajo valor. Sin embargo, presenta desafíos para detectar todas las ventas altas, lo que implica un riesgo de falsos negativos que puede limitar la cobertura total de auditoría predictiva.

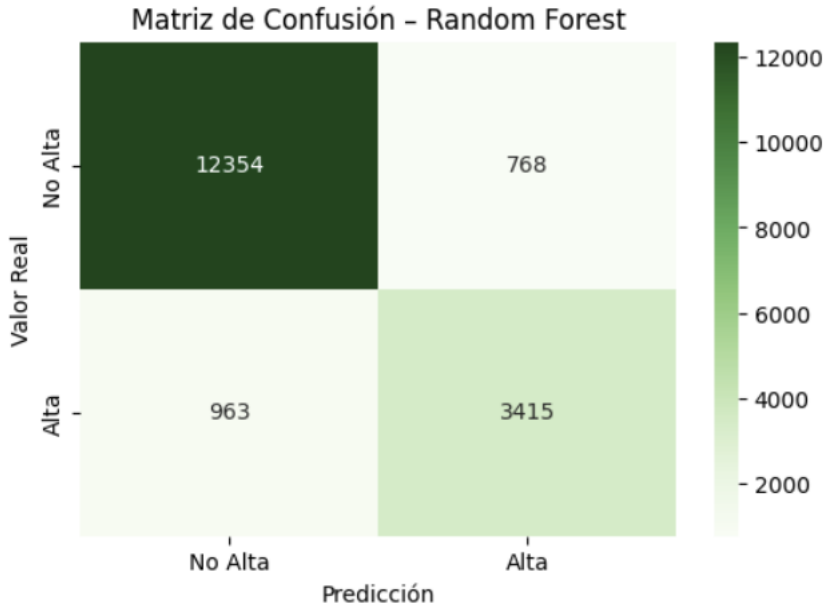
4.4.2.2. Random Forest

Complementando el análisis predictivo de ventas altas en las exportaciones de la empresa camaronera, se implementó un modelo de random forest que es un algoritmo de ensamble que combina múltiples árboles de decisión para mejorar la capacidad de predicción y reducir el sobreajuste característico de los modelos individuales.

4.4.2.2.1. Matriz de confusión

Figura 22:

Matriz de confusión - Random Forest - Exportaciones



La matriz de confusión presentada en la figura 22 evidencia el desempeño del modelo en la clasificación de transacciones como ventas no altas (0) y ventas altas (1):

Los verdaderos negativos hay 12,354 transacciones clasificadas correctamente como no altas.

En los falsos positivos hay 768 transacciones identificadas como altas sin serlas.

En los falsos negativos hay 963 transacciones de alto valor no detectadas por el modelo.

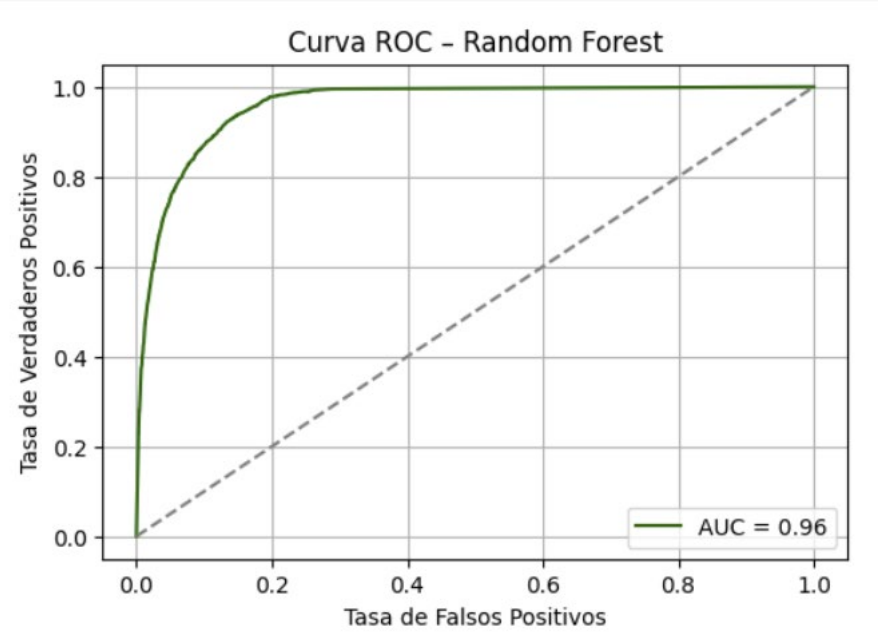
Dentro de los verdaderos positivos hay 3415 transacciones clasificadas correctamente como ventas altas.

Este resultado indica un alto nivel de acierto general, con un bajo número de falsos positivos y falsos negativos, lo que optimiza la identificación de operaciones relevantes para auditoría.

4.4.2.2.2. Curva ROC y AUC

Figura 23:

Curva ROC - Random Forest - Exportaciones



La curva ROC muestra una relación entre la tasa de verdaderos positivos (TPR o True positive rate en inglés) y la tasa de falsos positivos (FPR o false positive rate).

El modelo obtuvo un AUC = 0.96, lo que indica una capacidad discriminativa excelente, muy superior al desempeño aleatorio (AUC = 0.50).

4.4.2.2.3. Classification Report

Figura 24:

Classification report - Random Forest - Exportaciones

	precision	recall	f1-score	support
0	0.928	0.941	0.935	13122.000
1	0.816	0.780	0.798	4378.000
accuracy	0.901	0.901	0.901	0.901
macro avg	0.872	0.861	0.866	17500.000
weighted avg	0.900	0.901	0.900	17500.000

El informe de métricas muestra los siguientes resultados:

Clase 0 – Ventas no altas

Precisión = 0.928 y recall = 0.941, lo que indica que el modelo identifica con una gran confiabilidad las operaciones de bajo valor, disminuyendo la carga de revisión innecesaria.

F1-score es igual a0.935, confirmando un equilibrio sólido entre precisión y sensibilidad.

Clase 1 – Ventas altas

Precisión igual a 0.816, indicando que más del 81% de las predicciones de ventas altas son correctas. El resultado del recall de 0.780 implica que el modelo logra detectar el 78% de las transasacciones de alto valor, minimizando falsos negativos respecto al modelo de regresión logísitca.

F1-score es igual a 0.798 lo que refleja un desempeño perseverante y mejorado en comparación con el modelo anterior.

4.4.2.2.4. Métricas globales

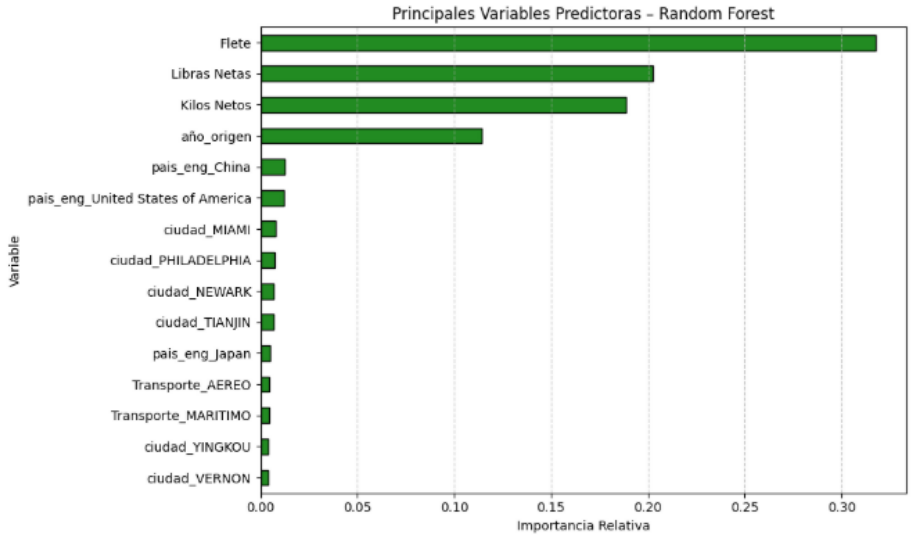
Accuracy de 0.901 lo que significa que el modelos clasifica correctamente el 90.1% de las transacciones.

Un macro avg de 0.866 lo que indica que tiene un promedio balanceado de desempeño en ambas clases.

Un weighted avg de 0.900 tiene un promedio ponderado por cantidad de registros, lo que confirma la confiabilidad global del modelo.

Figura 25:

Gráfico de variables importantes - Random Forest - Exportaciones



El presente análisis de importancia relativa evidencia que las variables cuantitativas como flete, libras y kilos netos tienen una fuerte influencia en la predicción de ventas altas, ya que operaciones de mayor valor económico y los volúmenes de carga suelen implicar costos logísticos más altos.

Año de origen aparece como variable relevante, indicando que existen tendencias temporales en los montos de las ventas, lo cual puede reflejar cambios en precios internacionales, la demanda de producto o políticas comerciales.

Entre las variables geográficas están los países de China y Estados Unidos como los destinos más influyentes. Por un lado, China muestra una relevancia alta por su alto volumen de

exportación, aunque en el modelo global algunas de estas operaciones pueden no representar ventas unitarias altas. Por otro lado, Estados Unidos, aunque con menor frecuencia, suele estar asociado a operaciones de alto valor unitario, lo que explica su aparición entre las variables relevantes.

Finalmente, este análisis muestra que el modelo random forest prioriza variables cuantitativas como determinantes de ventas altas, mientras que factores geográficos y logísticos tienen menor influencia relativa.

Para la auditoría financiera moderna, esta información es valiosa porque permite identificar los factores clave que concentran riesgo e impacto económico y concentrar la revisión de operaciones de alto volumen y costo logístico, optimizando recursos y aumentando la cobertura de riesgos significativos.

Conclusión

Considerando los resultados obtenidos, el modelo random forest se presenta como la herramienta más adecuada para la auditoría predictiva de exportaciones, ya que entre las ventajas están la maximización de la detección de ventas altas, reduciendo el riesgo de omitir transacciones críticas.

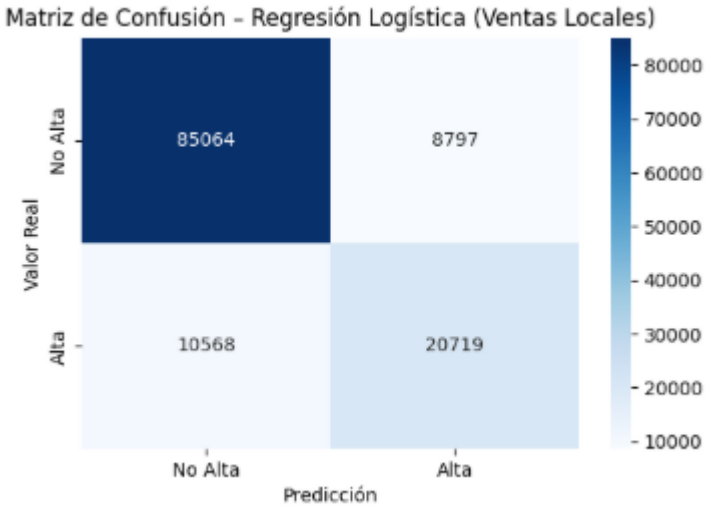
También permite concentrar la revisión analítica en operaciones de mayor impacto económico, también, optimiza recursos de auditoría al filtrar con alta precisión las operaciones menos relevantes.

Sin embargo, la regresión logística mantiene valor como modelo base para documentación y análisis interpretativo, complementando al enfoque predictivo de random forest dentro de un marco de auditoría financiera moderna basada en riesgos.

4.4.2.3. Regresión logística (mercado local)

Figura 26:

Matriz de Confusión - Regresión Logística | Ventas Locales



La matriz de confusión muestra la distribución de aciertos y errores del modelo para las ventas locales.

Muestra 85,064 verdaderos negativos, 8,797 falsos positivos, 10,568 falsos negativos y 20,719 verdaderos positivos.

Analizando el modelo distingue con mayor precisión las ventas no altas de clase 0 que las ventas altas de clase 1.

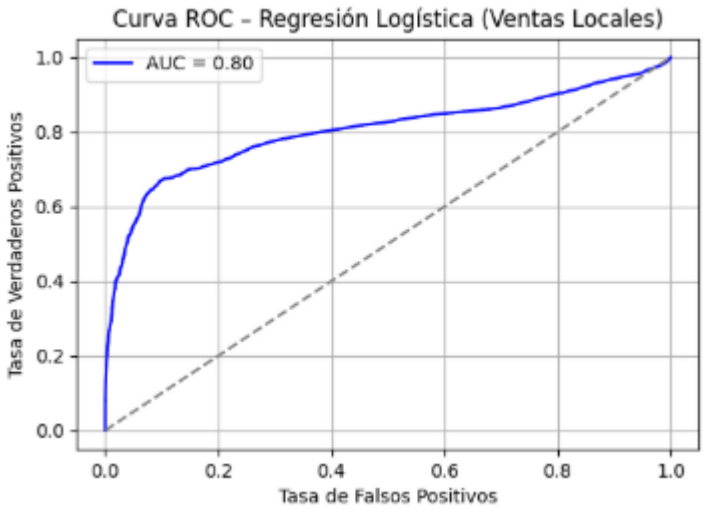
La cantidad de falsos negativos es relevante, lo que implica que aún hay ventas altas que no son identificadas correctamente, afectando la cobertura de control en auditoría.

Sin embargo, el número de verdaderos positivos indica que el modelo es capaz de priorizar un conjunto considerable de ventas de alto valor para análisis de riesgos.

4.4.2.3.1. Curva ROC y AUC (mercado local)

Figura 27:

Curva ROC y AUC | Regresión Logística | Ventas locales



La curva ROC refleja la capacidad del modelo para discriminar entre ventas altas y no altas en las ventas locales.

Según la figura 27 el área bajo la curva (AUC) es 0.80, lo cual indica un buen poder discriminativo; los valores de AUC cercanos a 1.0 reflejan excelente desempeño, mientras que un valor de 0.50 implicaría un modelo sin capacidad predictiva.

El modelo es confiable para auditoría predictiva, ya que logra balancear falsos positivos y falsos negativos.

4.4.2.3.2. Classification report (mercado local)

Figura 28:

Classification Report | Regresión Logística | Ventas Locales

	precision	recall	f1-score	support
0	0.889	0.906	0.898	93861.000
1	0.702	0.662	0.682	31287.000
accuracy	0.845	0.845	0.845	0.845
macro avg	0.796	0.784	0.790	125148.000
weighted avg	0.843	0.845	0.844	125148.000

En la clase 0, clasificada como ventas no altas se obtuvo una precisión de 0.889 lo que indica que la mayoría de las ventas predichas son realmente bajas. El modelo también tiene un recall de 0.906 afirmando que se identifican correctamente la mayor parte las ventas no altas.

El F1-score de 0.898 indica un alto equilibrio entre precisión y cobertura para esta clase.

En la clase 1, clasificada como ventas altas tiene una precisión de 0.702 lo que indica que siete de cada diez ventas clasificadas como altas realmente lo son; el recall de 0.662 indica que el modelo captura aproximadamente dos tercios de las ventas latas reales.

El F1-score de 0.682 indica un desempeño moderado, afectado principalmente por los falsos negativos.

En métricas globales el accuracy de 0.845 significa que el modelo acierto en 84.5% de los casos, lo cual es un valor aceptable en un contexto de análisis de transacciones.

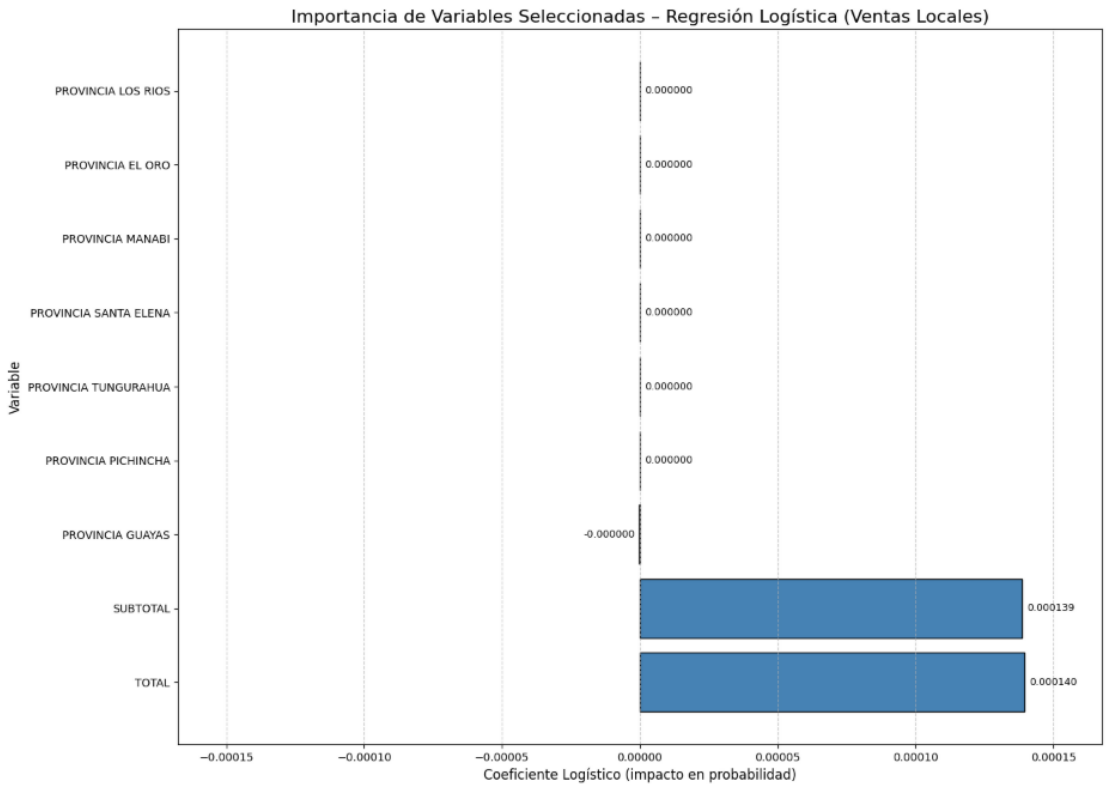
El macro avg F1 es de 0.790 indica un promedio balanceado entre ambas clases (refiriéndonos a la clase 1 y clase 0), muestra un desempeño consistente.

El weighted avg F1de 0.844 ajustado por la cantidad de datos, indica un buen comportamiento.

4.4.2.3.3. Importancia de variables

Figura 29:

Análisis de importancia de variables | Regresión logísitca | Ventas locales



La figura 29 indica el análisis de la importancia de variables mediante regresión logística permitió identificar los factores que inciden con mayor fuerza en la clasificación de una factura como venta alta. El modelo permite priorizar la revisión da facturas de mayor valor, confirmando que el valor monetario (Subtotal y Total) de las transacciones es el principal predictor de ventas altas, optimizando el uso de recursos en controles basados en riesgos.

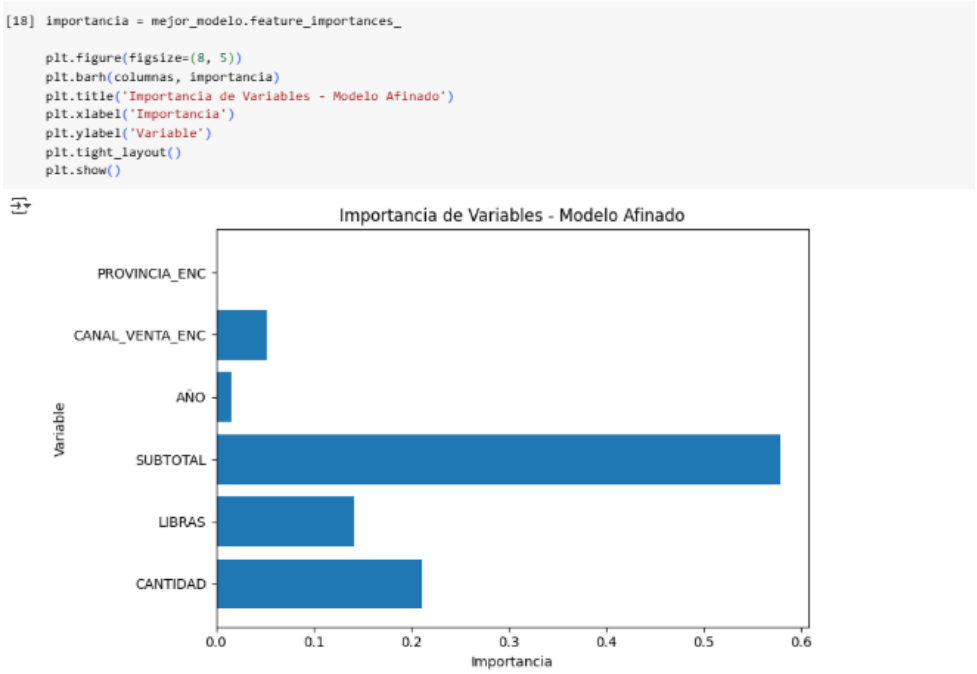
Las variables categóricas (provincias) cumplen un rol secundario, indicando que la ubicación geográfica del cliente no es factor determinante, pero aportan contexto para la segmentación de clientes y regiones.

4.4.2.4. Random forest (Ventas locales)

4.4.2.4.1. Importancia de variables – Modelo de random forest afinado – Ventas locales

Figura 30:

Importancia de Variables | Modelo afinado random forest | Ventas locales



El gráfico de barras muestra la importancia relativa de las variables predictoras en el modelo random forest afinado aplicado a las ventas locales de la empresa acuícola. Esta métrica fue calculada mediante el criterio de impureza promedio, que permite identificar qué atributos aportan más a la predicción del total facturado.

El subtotal se posiciona como la variable más determinante del modelo, lo cual es coherente desde el punto de vista contable y financiero al representar la base imponible de cada transacción, tiene una relación casi lineal con el total facturado, ya que tiene un reflejo directo al valor de los productos.

La cantidad y libras presentan una importancia considerable, aunque menor ya que va de 0.2 a 0.15. El peso evidencia que las dimensiones físicas de la venta complementan al subtotal al captar patrones relacionados con el volumen de la transacción.

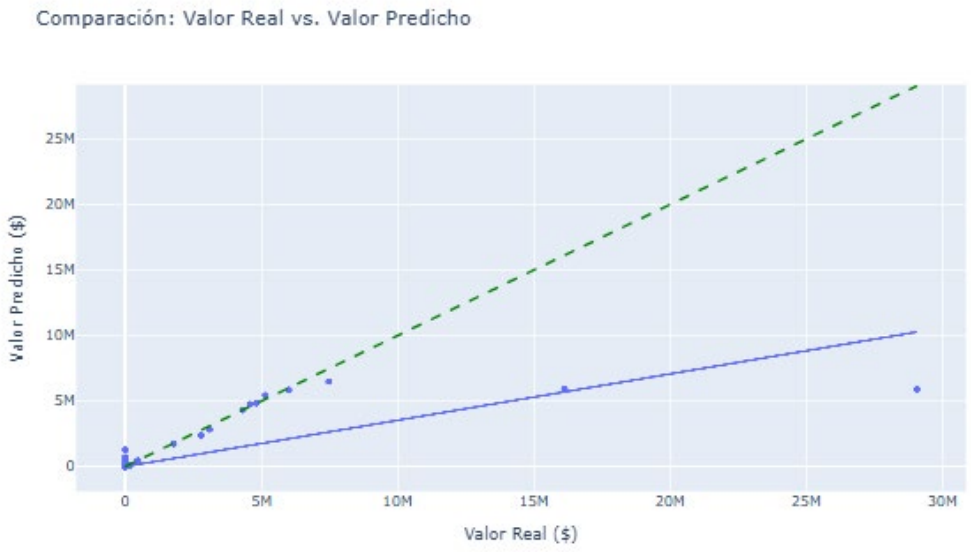
Esto es relevante en auditoría ya que la unidad vendida puede tener variaciones significativas en peso o presentación.

Las variables categóricas como canal de venta y provincia, aunque relevantes en un contexto comercial, muestran una contribución marginal al modelo.

4.4.2.4.2. Valor real vs. Valor predicho

Figura 31:

Gráfico de comparación: Valor Real vs. Valor Predicho



El gráfico de dispersión evalúa la capacidad predictiva del modelo ajustado comparando los valores reales contra los valores estimados por el modelo.

La línea verde discontinua representa el escenario ideal en el que cada valor predicho coincide exactamente con su valor real (y = x). Cualquier punto que se encuentre sobre esta línea indica una predicción perfecta.

La línea azul inferior a la línea verde sugiere que el modelo tiende a subestimar los valores más altos de facturación. Esto es particularmente visible en las facturas con montos superiores a los 10 millones, donde las predicciones se alejan de la diagonal perfecta.

La distribución de puntos en la zona de bajas ventas, el modelo muestra mayor densidad de aciertos, con dispersión reducida.

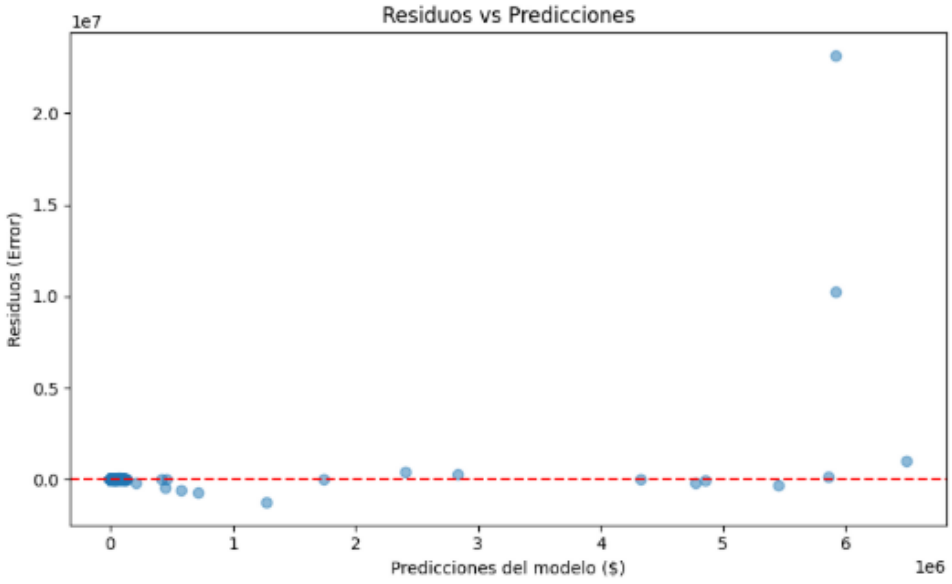
En valores altos, la dispersión aumenta, lo que refleja mayor error relativo en transacciones de montos altos.

En el contexto de auditoría, el hecho de que el modelo pierda exactitud en montos altos es relevante ya que las transacciones de gran valor suelen tener mayor riesgo e impacto en los estados financieros, por lo que este hallazgo justifica la aplicación de una revisión manual o modelos específicos para ese segmento.

4.4.2.4.3. Residuos vs. predicciones

Figura 32:

Gráfico de dispersión residuos vs. predicciones



La figura 32 presenta en el eje horizontal las predicciones del modelo en dólares y en el eje vertical los residuos o errores de predicción, calculados como la diferencia entre el valor real y el valor estimado (real – predicho). La línea roja discontinua en “y” = 0 indica el punto donde el error es nulo.

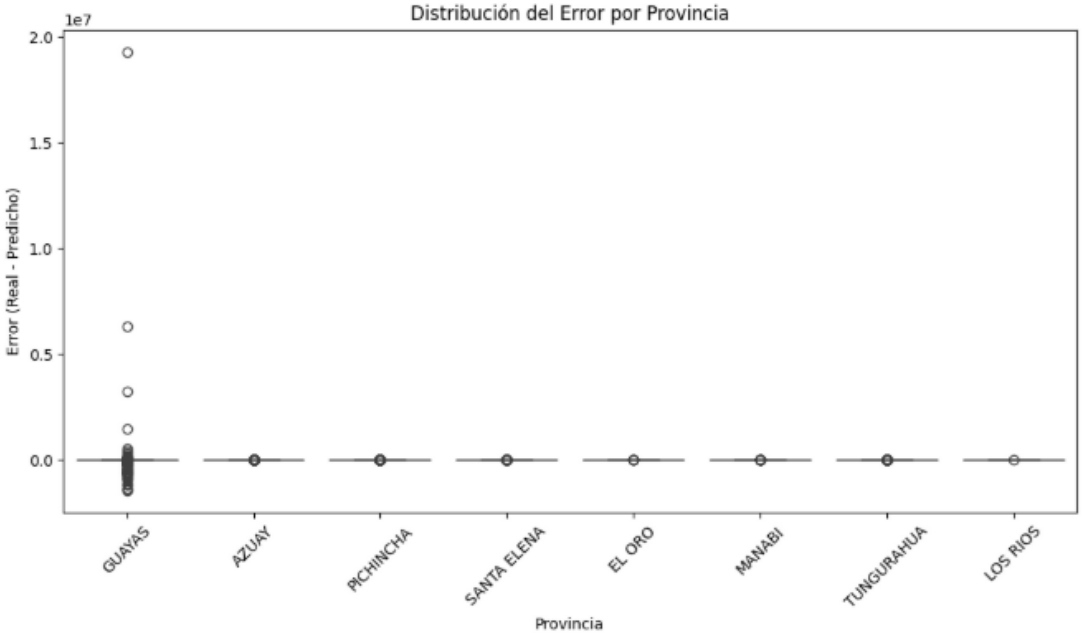
En el gráfico se observan puntos aislados muy por encima del resto, con residuos superiores a 20 millones en algunos casos. Estos casos extremos corresponden probablemente a facturas de alto valor en las que el modelo subestimó fuertemente el total. Este patrón coincide con lo observado en la comparación del valor real vs. valor predicho.

En auditoría, estos outliers no son simples errores de modelo, porque representan alertas sobre transacciones que se desvían significativamente de los patrones esperados, por lo que pueden ser priorizadas en procesos de revisión detallada. En control interno, estos casos podrían señalar desde errores de registro hasta prácticas comerciales atípicas o potenciales fraudes.

4.4.2.4.4. Boxplot de distribución de errores ventas locales

Figura 33:

Distribución de errores por provincia



El diagrama de caja como comportamiento la mayoría de las provincias exhiben errores cercanos a cero, con distribución relativamente simétrica y sin presencia significativa de valores extremos, lo que sugiere que el modelo mantiene un desempeño estable en gran parte del territorio.

La provincia del guayas presenta una dispersión de errores superior al resto, con varios outliers que alcanzan valores cercanos a los 20 millones. Este comportamiento puede estar vinculado a alta concentración de operaciones comerciales en esta provincia, lo que incrementa la probabilidad de transacciones atípicas de alto valor.

En provincias como Azuay, Pichincha, Santa Elena, El Oro, Manabí, Tungurahua y Los Ríos, los errores se mantienen estrechamente agrupados, lo que indica que en mercados de menor volumen el modulo predice con mayor exactitud relativa.

4.5. Fase 5: Evaluación

Antes de la fase de despliegue, la evaluación confirma si el modelo cumple un desempeño técnico sostenido, una utilidad real para auditoría, una calidad y consistencia de datos, y por último que tenga trazabilidad, ética y aceptación. La siguiente tabla resume qué se revisa, cómo se mide y quién decide el "Go/No-Go" (Autorización de despliegue / Detener y corregir).

Tabla 2

Evaluación del modelo

Bloque	Objetivo	Comprobación	Indicador	Umbral	Responsable	Decisión
1. Calidad de aciertos	Ver si el modelo distingue bien entre "normal" y "riesgoso"	Probar con datos de distintos periodos (sin volver a ajustar el modelo) y comparar contra el resultado real	Puntuación de diferenciación (0–1): más cerca de 1 = mejor	≥ 0,80	Tesista	Go si ≥ 0,80
2. Errores críticos	Controlar falsos avisos y casos importantes que se escapan	Contar aciertos y errores en una tabla simple (aciertos, falsos avisos, escapados)	De cada 100 avisos, cuántos fueron ciertos; de cada 100 importantes, cuántos detectó	Avisos ciertos ≥ 70%; Detectados ≥ 65%	Tesista – Dep. de auditoría	Go si cumple
3. Utilidad del auditor	Que ahorre tiempo revisando primero lo más relevante	Ordenar casos por "riesgo" y evaluar los primeros que se revisan	Aciertos en la lista priorizada (ej.: en los 200 primeros casos)	≥ 75% de aciertos en esa lista	Dep. de auditoría	Go si gana eficiencia

4. Estabilidad por segmentos	Que funcione parecido en años, canales y destinos distintos	Repetir las mediciones por año, canal, provincia/país, vendedor	Diferencia entre grupos	Variación ≤ 10 puntos	Tesista	Go si estable
5. Datos y ETL	Obtener datos completos, tipos correctos y totales que cuadran	Validar tipos, nulos y sumas por factura (en ventas locales)	Descuadres por factura y nulos críticos	0% descuadres; <1% nulos críticos	Sistemas + Contabilidad	Go si no hay hallazgos
6. Privacidad y accesos	Proteger datos sensibles y dar permisos correctos	Revisar máscaras de datos y roles de acceso	Cero hallazgos de acceso indebido	0 hallazgos	Sistemas + legal	Go si cumple
7. Decisión final	Autorizar uso o pedir mejoras	Revisión del comité con todo lo anterior	Realizar acta con “Ir” o “Esperar”	≥ 80% de ítems “Cumple” y sin críticos	Comité (Auditoría., sistemas, Gerencia)	Go / Hold

4.6. Fase 6: Despliegue

La fase final de despliegue convierte el prototipo en una solución útil y controlada para auditoría. Aquí pasamos del "modelo que funciona" a "proceso que aporta valor" con roles, entregables, controles y un arranque progresivo que va desde el piloto, la puesta en marcha y la operación con monitoreo y mejoras.

Tabla 3

Despliegue del modelo

Paso que seguir	Descripción de operación	Responsable	Entregable	Herramientas	Plazo
1. Piloto en colab	Ejecutar notebooks con dataset 2021–2024 (muestras representativas). Registrar métricas base (ROC/AUC, F1, recall clase "alta").	Tesista + Dep. sistemas	Notebook reproducible, seed fijada, métricas ≥ a línea base definida.	Google Colab + Google Drive	1–2 semanas
2. Validación Funcional	El departamento de auditoría revisa top-N transacciones "sospechosas/altas" marcadas por el modelo y documenta hallazgos.	Dep. de auditoría	Hoja de validación con % de aciertos útiles y feedback de reglas.	Notebook de Google Colab + Checklist	1 semana
3. Estandarizar datos	Realizar un pipeline de limpieza/estandarización (fechas, montos, agrupación por factura en ventas locales).	Sistemas + Contabilidad	Script ETL validado; pruebas sobre muestras con resultados consistentes.	Python/ETL	1 semana

4. Afinamiento	Re-entrenar datos. Revisar desbalance y costo-error.	Tesista	Mejora en recall de clase de interés sin degradar precision > umbral.	Colab/Sklearn/XGBoost	1 semana
5. Aprobación interna	El comité conformado por Auditoría, Contabilidad, TI y Gerencia valida que el modelo complementa revisión basada en riesgos.	Comité	Acta de aprobación y versión “v1.0” congelada.	Acta - Checklist	1 reunión
6. Escalamiento a dashboards	Publicar salidas en dashboard con listas priorizadas, filtros (año, canal, país/provincia, vendedor, etc.).	Sistemas + auditoría	Dashboard operativo, actualización mensual/semanal automatizada.	Power BI	1–2 semanas
7. Monitoreo continuo	Registro de métricas en producción (drift, % de alertas útiles, tiempo de revisión).	Auditoría + sistemas	Reporte mensual de comportamiento y recalibración trimestral.	Power BI	Mensual / Trimestral

4.6.1. Costos referenciales mínimos para el despliegue

Tabla 4

Costos de implementación

Concepto	Opción mínima viable	Recursos	Costo referencial
Almacenamiento de dato	Drive compartido básico de 100 – 200 GB	Datos históricos 2021-2024 + outputs y backups comprimidos	$10 / mes
Cómputo	Google colab Free / Pro para entrenamiento liviano	Datasets ya preprocesados, sesiones no intensivas	$10 / mes por usuario
BI / Dashboard	Looker Studio Free o 1 – 2 licencias Pro como alternativa	1 autor + 2 observadores gerenciales	$10 a $20 / mes por usuario
Control de versiones	Git repo privado	1 – 3 colaboradores	$4 / mes
Capacitación	Taller interno auditoría-ML (8–12 h)	6–10 personas, enfoque uso del dashboard y lectura de señales	$150 - $450 único pago
Mantenimiento	4–8 h/mes para ETL, recalibración trimestral	1 analista de datos + 1 auditor líder	$80 - $00 / mes

CAPÍTULO 5: ASPECTOS RELEVANTES DE LA PROPUESTA Y CONCLUSIONES

Partiendo de la problemática identificada en la fase de planteamiento, relacionada con la dificultad para detectar patrones irregulares en grades volúmenes de datos, se trabajó con la metodología CRISP-DM junto con algoritmos de aprendizaje supervisado y no supervisado. Esto permitió el análisis de data histórica de exportaciones y ventas locales de una empresa exportadora del sector acuícola, con el propósito de optimizar los procesos de auditoría financiera y detectar de manera temprana inconsistencias.

Los resultados obtenidos en la etapa de análisis revelaron patrones y comportamientos diferenciados tanto en mercado local como en el mercado de exportación.

En el mercado de exportaciones, los modelos no supervisados con DBSCAN e isolation forest permitieron identificar agrupaciones y valores atípicos en variables como el FOB, las libras netas y el CFR, dejando en evidencia transacciones que, por la magnitud de datos, podrían requerir una revisión. Las anomalías fueron más frecuentes en operaciones de altos volumen y en destinos específicos, lo que plantea la necesidad de controles diferenciados por país y tipo de producto.

Por otro lado, en el mercado local, el análisis con DBSCAN sobre variables agrupadas por factura, precio promedio y total de la transacción identificó concentraciones de ventas coherentes con el patrón histórico y un conjunto reducido de operaciones fuera de rango, que podrían explicarse como quizás errores de registro, ventas que no tengan que ver con el giro del negocio o descuentos no procesados. Este hallazgo tiene importancia para la auditoría, ya que permite enfocar la revisión en un subconjunto pequeño de transacciones que son significativas.

Uno de los aspectos relevantes de este proyecto es la formulación de un producto mínimo viable (MVP) de análisis predictivo para la auditoría financiera en la empresa exportadora. El modelo central propuesto es el Random Forest, ya que tiene la capacidad de manejar relaciones no lineales y estabilidad en los resultados. A través de este modelo fue posible clasificar las transacciones comerciales en dos categorías: ventas altas y ventas no altas, definiéndose la variable objetivo en función del valor CFR por el lado de las exportaciones y por el lado de ventas locales el subtotal. Esta definición es particularmente relevante para la auditoría, ya que las ventas de alto valor representan un mayor riesgo financiero y, por lo tanto, merecen prioridad en los procesos de control y revisión.

En la construcción del MVP no se aplicaron técnicas de balanceo de clases ni validación cruzada, lo que constituye una limitación en la detección de casos minoritarios (ventas altas). No obstante, se reconoce que en futuras iteraciones del modelo estos elementos deberán incorporarse para mejorar la capacidad de generalización y la detección de operaciones de bajo volumen pero alto impacto en los riesgos financieros.

Es importante resaltar que este MVP constituye un ejercicio académico, desarrollado y ejecutado en un entorno de Google Colab, con los datos históricos de exportaciones y ventas locales. Para su adopción en un entorno empresarial, será necesario evaluar el despliegue en otras plataformas de integración y producción, que garanticen seguridad, y facilidad de uso por parte de los equipos contables, sistemas y de auditoría.

Los modelos supervisados, como random forest y regresión logística demostraron una capacidad predictiva muy consistente, ya que métricas de F1-score y matrices de confusión validan la utilidad para clasificar transacciones según criterios de riesgo.

Particularmente random forest mostró una mayor estabilidad y precisión, lo que lo convierte en un candidato ideal para integrarse en un sistema de monitoreo continuo.

La preparación y calidad de la data es un factor determinante para el éxito del sistema. Las etapas de limpieza, estandarización y consolidación, incluyendo la agrupación de registros por número de factura o códigos de embarque en las exportaciones, fueron parte esencial para optimizar el rendimiento de los modelos y garantizar la fiabilidad de los resultados. Esta metodología mejora la interpretabilidad de los hallazgos y reduce la probabilidad de falsos positivos o negativos.

En términos de recursos y roles, la propuesta se integra de manera flexible al entorno organizacional existente (área contable, auditoría, sistemas, gerencias). El uso de herramientas en la nube como Google colab y Google drive permite minimizar costos de infraestructura y facilita la colaboración entre departamentos, posibilitando que el área contable, auditoría y los usuarios de comercio exterior y ventas locales trabajen de forma coordinada en el análisis de datos. La solución también incluye capacitar al personal para la interpretación de resultados y la generación de reportes a gerencia, para así fortalecer la toma de decisiones estratégicas.

En conclusión, implementar este sistema de análisis predictivo representa un avance significativo hacia un modelo de auditoría preventivo y basado en evidencias que encontramos en datos. La adopción de este incrementa la capacidad de la organización para anticipar riesgos financieros, detectar irregularidades y reforzar el control interno de la compañía, mejorando las prácticas de auditoría dentro de la misma; la propuesta puede adaptarse a diferentes volúmenes de datos y contextos empresariales, lo que garantiza su aplicación a largo plazo.

5.1. Consideraciones finales

Si bien el sistema propuesto ofrece resultados significativos y aporta valor en la detección de anomalías financieras, es necesario reconocer ciertos riesgos y limitaciones asociados a su implementación. En primer lugar, los modelos de ML pueden volverse obsoletos con el tiempo debido a cambios en las dinámicas comerciales, la variación en los mercados internacionales o la introducción de nuevos productos, lo que requiere un mantenimiento y reentrenamiento periódico. En segundo lugar, la calidad de los resultados depende en gran medida de la integridad y consistencia de los datos, registros incompletos, duplicados o con errores pueden afectar la capacidad predictiva del modelo.

Finalmente, es posible enfrentar resistencia al cambio por parte del personal de la empresa, al introducir una herramienta tecnológica que transforma prácticas tradicionales de control y de auditoría hacia un enfoque más automatizado y preventivo.

Por otro lado, es fundamental destacar cómo el modelo se conecta directamente con la auditoría financiera real. En la práctica, el sistema no sustituye al auditor, sino que actúa como una herramienta de apoyo que prioriza y clasifica la información. Las transacciones que el modelo identifica como atípicas o sospechosas se convierten en los primeros casos que serán revisados por el auditor, permitiendo un uso más eficiente de su tiempo y enfocando la atención en aquellas operaciones con mayor probabilidad de riesgo. Este proceso agiliza la detección de inconsistencias, reduce la revisión manual de grandes volúmenes de datos y ayuda en la transparencia en los informes financieros, convirtiendo la auditoría en un proceso más preventivo que correctivo.

Para concluir, la incorporación de este sistema predictivo no solamente moderniza los procesos de auditoría en la empresa acuícola, sino que también abre el camino hacia una cultura

organizacional más orientada al análisis de datos, la eficiencia y la gestión proactiva de riesgos, alineándose con las mejores prácticas en auditoría financiera.

BIBLIOGRAFIA

(CNA), C. N. (2024). Revista Aquacultura. *Aquacultura, la voz oficial del sector*.

(FAO), O. d. (2022). *The State of World Fisheries and Aquaculture 2022.* Roma: FAO. https://doi.org/10.4060/cc0461en

Arens, A. A., Elder, R. J., Beasley, M. S., & Hogan, C. E. (2018). *Auditing and Assurance Services.* England: Pearson.

Bartz-Beielstein, T. (2024). *Supervised Learning: Classification and Regression.* Springer. https://doi.org/https://doi.org/10.1007/978-981-99-7007-0_2

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning.* Springer. https://doi.org/https://doi.org/10.1007/978-1-0716-1418-1

Levy, M. (2021). *Dataquest*. Dataquest. https://www.dataquest.io/tutorial/telling-data-stories-with-python-using-information-design/

Malheiro, L., Leocádio, D., & Reis, J. (2024). Artificial Intelligence in Auditing: A Conceptual Framework for Auditing Practices. *MDPI*, 1-2. https://doi.org/https://doi.org/10.3390/admsci14100238

Mariano, R. M., Pedro, C. I., & José, P. C. (2021). Utilidad del Deep Learning en la predicción del fracaso empresarial en el ámbito europeo. *Revista de Metodos Cuantitativos para la Economia y la Empresa*, 392-414. https://doi.org/10.46661/revmetodoscuanteconempresa.5172

Peñarreta-Angamarca, M. T., Torres-Palacios, M. M., & Moreno-Narváez, V. P. (2024). Efectividad de la auditoría financiera en la prevención del fraude en pequeñas y medianas

empresas. *Revista Multidisciplinaria Perspectivas Investigativas*, 26-35. https://doi.org/ https://doi.org/10.62574/rmpi.v4iespecial.106

Russel, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach (4th ed.)* (Cuarta ed.). Pearson.

Sun, T., & Vasarhelyi, M. A. (2018). Embracing Textual Data Analytics in Auditing. *The International Journal of Digital Accounting Research*, 19. https://doi.org/10.4192/1577-8517-v18_3

Suresh Rao, A., Vishnu, B. V., & Shaik, H. (2021). Role of Exploratory Data Analysis in Data Science. *2021 6th International Conference on Communication and Electronics Systems (ICCES).* Coimbatre, India: IEEE. https://doi.org/10.1109/ICCES51350.2021.9488986

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). *Big Data in Accounting: An Overview.* Accounting Horizons. https://doi.org/10.2308/acch-51071

Cruz Ramírez, E. S., Intriago Sánchez, N. A., & Landeta Púa, D. C. (2023). *Modelo para pronosticar la demanda de alimentos balanceados en el sector acuícola* [Escuela Superior Politécnica del Litoral]. https://www.dspace.espol.edu.ec/handle/123456789/60502

Galvez Ferrerira, J. D., Nieto Rodrigues, M. P., & Rocha Ruiz, C. A. (2022). Prediciendo el crimen en ciudades intermedias: un modelo de “machine learning” en Bucaramanga, Colombia. *URVIO Revista Latinoamericana de Estudios de Seguridad*. https://doi.org/https://doi.org/10.17141/urvio.34.2022.5395

Morán Quispe, P. M. (2022). *Relación entre algunos modelos estadísticos clásicos y modelos de RNA: Un análisis comparativo* [Universidad Nacional de Piura]. https://repositorio.unp.edu.pe/items/93c86d1d-91dd-45b6-b111-7a78225f01f8

Reyes Sarmiento, T. P. (2022). *Modelo de optimización de procesos bancarios o financieros para agilizar procedimientos relacionados mediante Business Intelligence*. http://dspace.ups.edu.ec/handle/123456789/23337