Maestría en

CIENCIA DE DATOS Y MÁQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

Trabajo Previo a la Obtención de Título de Magister en Ciencia de Datos y Máquinas de Aprendizaje

AUTOR/ES:

Andino Guerra Christian Adrián

Montenegro Changotasi Alex Alejandro

Tituaña Cevallos Ingrid Xiomara

Tufiño Salazar Cynthia Stefania

TUTOR/ES:

Karla Mora

Alejandro Cortés

TEMA:

Modelo Predictivo Dual para Analizar la Propensión Exportadora y el Volumen Exportado de Empresas Ecuatorianas

Quito - Ecuador

diciembre – 2025

REINVENTEMOS EL FUTURO

Certificación de Autoría

Nosotros, **Andino Guerra Christian Adrián, Montenegro Changotasi Alex Alejandro, Tituaña Cevallos Ingrid Xiomara y Tufiño Salazar Cynthia Stefania,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma
Andino Guerra Christian Adrián

Firma
Montenegro Changotasi Alex Alejandro

Firma
Tituaña Cevallos Ingrid Cevallos

Firma
Tufiño Salazar Cynthia Stefania

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Andino Guerra Christian Adrián, Montenegro Changotasi Alex Alejandro, Tituaña Cevallos Ingrid Xiomara y Tufiño Salazar Cynthia Stefania**, en calidad de autores del trabajo de investigación titulado ***Modelo Predictivo Dual para Analizar la Propensión Exportadora y el Volumen Exportado de Empresas Ecuatorianas***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, diciembre 2025

Firma
Andino Guerra Christian Adrián

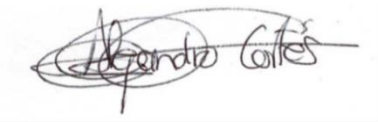
Firma
Montenegro Changotasi Alex Alejandro

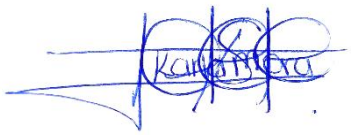
Firma
Tituaña Cevallos Ingrid Cevallos

Firma
Tufiño Salazar Cynthia Stefania

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés Director EIG y Karla Mora Coordinadora UIDE,** declaramos que: **Andino Guerra Christian Adrián, Montenegro Changotasi Alex Alejandro, Tituaña Cevallos Ingrid Xiomara y Tufiño Salazar Cynthia Stefania**, son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.





Alejandro Cortés López

Director de la

Maestría en Ciencia de Datos y Maquinas de Aprendizaje con Mención en Inteligencia Artificial

---

Karla Estefanía Mora Cajas

Coordinadora de la

Maestría en Ciencia de Datos y Maquinas de Aprendizaje con Mención en Inteligencia Artificial

Dedicatoria

A mis padres, Christian y Adriana, por el acompañamiento permanente, los consejos diarios y el ejemplo de constancia que han guiado cada etapa de mi formación académica y personal.

A mis hermanas, Adry y Bri, lo mejor de mi mundo, parte de mi motivación cotidiana.

Que este logro sea un punto de partida para muchos éxitos compartidos

míos y de ustedes; de ustedes y míos

Andino Guerra Christian Andino

A mi familia, que han estado conmigo de forma incondicional.

Montenegro Changotasi Alex Alejandro

A mi familia, mi raíz y motor de vida.

A mis padres, gracias por su amor incondicional, por sostenerme con su ejemplo y recordarme cada día que sí puedo. Con ustedes aprendí que el amor verdadero se demuestra con presencia, paciencia y constancia, y que nunca falla; son mi lugar seguro y mi bendición.

A mi hijo, mi luz y mi mayor bendición. Me inspiras a ser mejor, más paciente y valiente; por ti elijo avanzar, aprender y cuidar lo esencial.

A mis hermanas, gracias por caminar conmigo, por el apoyo, la complicidad y el amor que nos une. Su presencia me fortalece y me recuerda que no estoy sola.

A mi sobrina, gracias por tu alegría y ternura, por unirnos tanto y por llenar de vida nuestro hogar; haces que la familia se sienta más completa.

Tituaña Cevallos Ingrid Cevallos

A mi padre, porque las diferencias de hoy no borran el amor de siempre

A mi madre, soy muy afortunada por tenerte en mi vida

A mi hermana Genesis, mi eterna compañera

A mi sobrina que es mi mayor inspiración

Y a mi hermano Oli, te quiero mucho más de lo que puedo expresar

Gracias por tanto.

Tufiño Salazar Cynthia Stefania

Agradecimientos

A todas las personas que formaron parte de este logro, principalmente a nuestras familias que son quienes nos apoyan cada día y nos inspiran a seguir adelante.

A nuestros amigos que hacen nuestra vida más amena.

Y a los profesores que formaron parte de este camino, gracias por compartir su conocimiento con nosotros y ayudarnos a crecer tanto profesional como laboralmente.

Cynthia, Christian y Alejandro

Agradezco primero a Dios por su guía, su fuerza y las bendiciones que me han sostenido en cada paso. A mi familia, gracias por ser mi base y mi impulso: por su amor, su apoyo constante y por no soltarme en los días difíciles. Este logro también es suyo, porque lo construimos con fe, esfuerzo y unión. A la universidad y a mis docentes, gracias por su enseñanza, paciencia y acompañamiento durante este proceso de formación.

Tituaña Cevallos Ingrid Cevallos

Resumen

El presente estudio tiene como objetivo el desarrollo y evaluación de un sistema predictivo enfocado en el análisis del comportamiento exportador de las empresas ecuatorianas, utilizando técnicas de aprendizaje automático. En este sentido se estructuró un sistema con un enfoque dual que desagrega el proceso de exportación en dos etapas, primero se calcula la probabilidad de que una empresa exporte esta es la etapa de clasificación y por otro lado en la segunda etapa se estima el monto de exportación, la combinación de ambas fases brinda a nivel poblacional una predicción integrada del valor esperado de exportación.

Para la investigación se utilizó la base de datos de la Encuesta Estructural Empresarial 2023, en base a la literatura se seleccionaron variables de ventas, productividad y de estructura empresarial.

Para el modelo de clasificación se compararon los modelos basados en Random Forest y XGBoost, siendo el segundo el modelo que mostró un mejor desempeño con métricas que muestran robustez y precisión (ROC-AUC alcanzó un valor de 0.93). Mientras que, para la estimación del monto exportado fue el modelo basado en Random Forest Regressor el que mostró resultados más sólidos captando una mayor proporción de la variabilidad asociada al fenómeno de exportación.

Los resultados demuestran que este sistema es sólido metodológicamente y puede constituir una herramienta eficaz en la toma de decisiones orientadas a potenciar el sector exportador ecuatoriano.

Palabras claves: aprendizaje automático, exportaciones, clasificación, regresión, empresas.

Abstract

The present study aims to develop and evaluate a predictive system focused on the analysis of the export behavior of Ecuadorian firms using machine learning techniques. To this end, a dual-stage approach was implemented, which decomposes the export process into two components: first, the probability that a firm exports is estimated through a classification stage; second, the export amount is predicted through a regression stage. The combination of both phases provides an integrated population-level prediction of the expected export value. The analysis is based on data from the 2023 Structural Business Survey, from which variables related to sales, productivity, and firm structure were selected according to the relevant literature. For the classification task, Random Forest and XGBoost models were compared, with XGBoost exhibiting superior performance, achieving robust and accurate results, as reflected by a ROC-AUC value of 0.93. In contrast, for the estimation of the export amount, the Random Forest Regressor model produced more stable results, capturing a larger proportion of the variability associated with export activity.

The results demonstrate that the proposed system is methodologically robust and can serve as an effective tool for decision-making aimed at strengthening Ecuador's export sector.

Keywords: machine learning, exports, classification, regression, enterprises.

Índice General

Índice de Tablas

Lista de Figuras

Capítulo I

1. Introducción

La dinámica exportadora ecuatoriana ha cobrado una relevancia decisiva para comprender la competitividad de las empresas y, al mismo tiempo, para orientar políticas que permitan ampliar la base productiva del país; el comportamiento del sector empresarial frente a los mercados internacionales no solo responde a factores macroeconómicos; también se vincula con diferencias profundas entre firmas en cuanto a productividad, escala operativa, acumulación de capital y estrategias de inversión, la teoría moderna de comercio internacional basada en heterogeneidad de empresas, impulsada por trabajos de autores, abrió un debate sobre los mecanismos que explican por qué algunas organizaciones avanzan hacia la exportación mientras otras permanecen en el mercado interno, esa perspectiva permitió entender que participar en el comercio exterior exige superar umbrales mínimos de eficiencia y enfrentar costos específicos asociados al proceso exportador, lo que genera un patrón selectivo donde solo las firmas más capaces logran competir fuera del territorio nacional (Morales y Contreras, 2024).

En el país se observa un patrón donde la mayor parte del tejido empresarial continúa orientado al mercado local; sin embargo, convive con un conjunto más reducido de compañías que ha conseguido abrir espacios en el comercio internacional, cada una con niveles distintos de presencia, esa diversidad responde a contrastes entre sectores, diferencias territoriales y brechas en la capacidad de invertir en tecnología, lo que termina configurando un escenario en el que la posibilidad de exportar y la cantidad enviada al exterior se ven influidas por varios factores que pueden medirse, entre ellos se incluyen la dimensión de la empresa, su volumen de ventas, los gastos asociados al traslado de mercancías y la manera en que se organiza la producción (Paredes, Vinces, Castro, & Llerena, 2025).

Estas condiciones generan relaciones difíciles de capturar mediante enfoques lineales simples, ya que las variables interactúan de formas que no siempre se aprecian a primera vista; por esa razón se vuelve necesario emplear métodos que permitan reconocer comportamientos irregulares, variaciones pronunciadas y efectos combinados entre características económicas y operativas, dentro de ese marco, la ciencia de datos y los algoritmos de aprendizaje resultan especialmente útiles, dado que ofrecen una forma de representar escenarios con alta dispersión, presencia de incertidumbre y distribuciones donde un pequeño grupo concentra montos elevados, fenómeno frecuente en los valores exportados (Paredes, Vinces, Castro, & Llerena, 2025).

El diseño de un modelo predictivo dual, que combine por un lado un clasificador para estimar la probabilidad de que una empresa participe en actividades de exportación y, por otro lado, un modelo de regresión orientado a predecir el volumen exportado condicional a esa participación, permite abordar el fenómeno desde una perspectiva integral; la separación entre participación y monto responde a una distinción conceptual ampliamente reconocida en la literatura: la decisión de exportar constituye un comportamiento cualitativamente diferente al nivel de intensidad exportadora, este enfoque posibilita un análisis más fino de los determinantes y, además, mejora la precisión de las predicciones en escenarios donde la distribución del monto exportado muestra alta asimetría (Pinos, Cevallos, Abril, & Pauta, 2025)

El empleo de algoritmos como *RandomForest, XGBoost*, máquinas de vectores de soporte y otros regresores de carácter no lineal brinda la posibilidad de analizar vínculos complejos entre variables económicas y características propias de cada empresa; este trabajo se complementa con un tratamiento previo de los datos que ayuda a suavizar valores atípicos y a dar mayor estabilidad a las distribuciones mediante transformaciones como log1p o

recortes controlados de colas, a ello se incorpora el uso de los pesos muestrales de la Encuesta Estructural Empresarial del INEC, que se aplican tanto en la fase de entrenamiento como en la evaluación del desempeño, esta inclusión permite obtener resultados que reflejan de mejor manera la estructura real del tejido empresarial del país, aspecto fundamental cuando se busca fundamentar decisiones públicas u orientar estrategias en el ámbito productivo (Ferrouhi y Bouabdallaoui , 2025).

Por tal motivo, la comprensión de los resultados del modelo compone un tema fundamental, en este marco, herramientas como *Python* permiten identificar patrones que influyen de manera directa en la propensión exportadora de las empresas, ya sea favoreciendo o limitando su participación en mercados internacionales, lo que aporta mayor claridad al proceso metodológico y facilita que los resultados puedan ser utilizados por entidades públicas, gremios productivos y organismos encargados de la promoción comercial; a partir de ello, la estimación del valor esperado de las exportaciones a nivel empresarial, construida mediante la combinación entre la probabilidad de exportar y el monto proyectado, brinda solidez para diseñar esquemas de priorización sustentados en evidencia, orientados a aquellas firmas que presentan mejores condiciones para ampliar su presencia en el comercio exterior (Martínez & Cobos, 2025).

1.1 Formulación del Problema

El desempeño exportador de las empresas ecuatorianas se caracteriza por una marcada desigualdad entre aquellas que logran participar de manera sostenida en los mercados internacionales y un grupo mayoritario que permanece orientado exclusivamente al mercado interno, por lo tanto, esta diferencia no responde a una causa única ni a un patrón homogéneo, sino que se configura a partir de una combinación de factores económicos, productivos y

estructurales que influyen de manera distinta en cada firma; en la práctica, empresas que operan dentro de un mismo sector o incluso en una misma región muestran comportamientos muy disímiles frente al comercio exterior, lo que evidencia la presencia de una heterogeneidad empresarial que no siempre es capturada por los enfoques de análisis tradicionales (Aldrin, Aviles, Baque, & Muñiz, 2024).

A nivel institucional, el estudio del comercio exterior suele centrarse en indicadores agregados como el valor total exportado, la participación sectorial o el saldo de la balanza comercial, si bien estos indicadores resultan útiles para describir la evolución general del sector externo, presentan limitaciones cuando se busca comprender el comportamiento de las empresas como unidades individuales, en particular, este enfoque dificulta responder preguntas clave relacionadas con la probabilidad de que una empresa decida exportar y con el volumen que podría alcanzar una vez que participa en el comercio internacional, la ausencia de herramientas que permitan anticipar estos comportamientos restringe la capacidad de diseñar políticas públicas más focalizadas y reduce la efectividad de los programas de promoción exportadora (Morales y Contreras, 2024).

Por otro lado, la dificultad del análisis aumenta cuando se reconoce que la decisión de exportar y el nivel de ventas externas no obedecen a una misma lógica; la participación en mercados internacionales suele relacionarse con la capacidad de asumir costos fijos, cumplir exigencias normativas y establecer vínculos comerciales estables, mientras que el monto exportado se ve condicionado por elementos adicionales como la escala de producción, la eficiencia en los procesos, el acceso a capital y la forma en que se estructuran los costos logísticos, es decir, estas dimensiones no actúan de manera aislada, sino que se influyen mutuamente y dan lugar a relaciones complejas que rara vez siguen trayectorias lineales,

circunstancia que limita la capacidad de los modelos econométricos tradicionales para representarlas adecuadamente bajo supuestos rígidos (Pinos, Cevallos, Abril, & Pauta, 2025).

En el caso ecuatoriano, esta complejidad se ve reforzada por la diversidad del tejido productivo, es decir, el país presenta una estructura empresarial compuesta en su mayoría por pequeñas y medianas empresas, con niveles variables de productividad, acceso desigual a financiamiento y diferencias importantes en la adopción de tecnología, frente a esto, se suman las brechas territoriales que influyen en los costos de transporte, la cercanía a puertos y la disponibilidad de infraestructura, estas condiciones hacen que el comportamiento exportador esté marcado por una alta dispersión y por la presencia de valores extremos en los montos exportados, donde un número reducido de empresas concentra una parte significativa del total (Cruz & Brito, 2023).

Desde el punto de vista analítico, esta realidad plantea un desafío importante, cabe señalar que, los métodos tradicionales permiten identificar asociaciones promedio entre variables, aunque presentan dificultades para capturar interacciones complejas y comportamientos irregulares, en particular, estos enfoques suelen ser sensibles a valores atípicos y presentan limitaciones cuando se trabaja con distribuciones sesgadas, situación frecuente en los datos de exportaciones, como resultado, las estimaciones obtenidas pueden no reflejar adecuadamente la variabilidad observada en el comportamiento de las empresas, lo que reduce su utilidad para fines predictivos y de planificación.

Frente a este escenario, surge la necesidad de contar con herramientas analíticas capaces de abordar el problema desde una perspectiva distinta, por lo mismo, la ciencia de datos y las técnicas de aprendizaje automático ofrecen un marco metodológico adecuado para analizar grandes volúmenes de información, modelar relaciones no lineales y manejar

interacciones complejas entre variables, no obstante, su aplicación al estudio del comercio exterior empresarial en el contexto ecuatoriano todavía es limitada, especialmente cuando se busca integrar de manera explícita la estimación de la propensión exportadora con el análisis del volumen exportado (Garzón & Díaz, 2023)

Un aspecto adicional que incrementa la complejidad del planteamiento del problema se vincula con la representatividad de la información empleada; las encuestas empresariales, entre ellas la Encuesta Estructural Empresarial del INEC, incorporan pesos muestrales concebidos para reproducir la composición real del universo empresarial, aunque en muchos casos dichos pesos no se incorporan de forma sistemática en los modelos predictivos, situación que puede dar lugar a estimaciones sesgadas y restringir la validez de los resultados cuando se interpretan a nivel poblacional.

En este contexto, el problema central de la investigación se configura alrededor de la dificultad para anticipar, con precisión y representatividad, qué empresas ecuatorianas tienen mayor probabilidad de participar en mercados internacionales y qué volumen de exportaciones podrían alcanzar; esta dificultad se traduce en limitaciones prácticas para la formulación de políticas públicas, la asignación eficiente de recursos y la toma de decisiones estratégicas por parte de las propias empresas, por lo que, la falta de un dirección que tome en cuenta la predicción integrado impide aprovechar plenamente la información disponible y reduce el potencial de los análisis basados en datos.

Por lo expuesto, el problema incorpora además la necesidad de distinguir con claridad entre la decisión de exportar y el nivel de intensidad exportadora; abordar ambos fenómenos como si fueran una sola dimensión puede ocultar procesos relevantes y dar lugar a lecturas parciales del comportamiento empresarial, de ahí que resulte pertinente adoptar un enfoque

que permita analizar estas dos etapas por separado, tomando en cuenta que se encuentran influenciadas por factores que no son completamente coincidentes y que demandan herramientas analíticas diferenciadas.

Desde esta perspectiva, la formulación del problema se orienta a la construcción de un modelo predictivo dual capaz de estimar, de manera diferenciada y complementaria, tanto la probabilidad de que las empresas ecuatorianas exporten como el volumen esperado de sus ventas externas, mediante el uso de técnicas de ciencia de datos que permitan capturar relaciones complejas, incorporar la representatividad poblacional y generar resultados comprensibles, es evidente que, abordar este desafío requiere no solo la elección de métodos adecuados, sino también la definición de un esquema metodológico alineado con las particularidades del tejido empresarial ecuatoriano y con las necesidades operativas de quienes toman decisiones en el ámbito del comercio exterior.

1.1.1 Sistematización del problema

A partir de la formulación planteada, el problema general se organiza en interrogantes que orientan la investigación:

1. ¿Qué características empresariales influyen con mayor fuerza en la probabilidad de que una firma ecuatoriana participe en actividades de exportación?
2. ¿De qué manera se relacionan variables como ventas, activos, inversión y costos de transporte con la magnitud del monto exportado por las empresas que ya operan en mercados internacionales?
3. ¿Cómo pueden los métodos de ciencia de datos captar interacciones no lineales y patrones complejos que afectan tanto la propensión a exportar como el volumen enviado al exterior?

4. ¿Qué tan preciso puede ser un modelo dual que combine una etapa de clasificación para estimar la participación y otra de regresión para anticipar el monto exportado?

5. ¿En qué medida la incorporación de pesos muestrales mejora la representatividad de las predicciones y permite obtener resultados más alineados con la estructura real del universo empresarial ecuatoriano?

6. ¿Cómo pueden utilizarse las estimaciones obtenidas para orientar decisiones de política pública, priorización sectorial y estrategias empresariales?

1.2 Importancia y Justificación del Estudio

La importancia de esta investigación se fundamenta en la necesidad de comprender con mayor profundidad el comportamiento exportador de las empresas ecuatorianas desde una perspectiva que vaya más allá de la descripción agregada de cifras comerciales; en el contexto nacional, la exportación suele analizarse a partir de volúmenes totales, balanza comercial o participación sectorial; sin embargo, estas aproximaciones dejan en segundo plano a la empresa como unidad de análisis y dificultan la identificación de los factores que explican por qué ciertas firmas logran integrarse al comercio internacional mientras otras permanecen orientadas exclusivamente al mercado interno. Este estudio adquiere relevancia al proponer un enfoque que permite observar esas diferencias desde el nivel microeconómico, considerando la heterogeneidad empresarial que caracteriza al tejido productivo del país (Morales y Contreras, 2024).

Dentro de esta perspectiva, la justificación del estudio también se vincula con la complejidad inherente al proceso exportador, en efecto, la decisión de exportar y el volumen alcanzado no dependen de un único elemento, sino de la interacción de múltiples factores

económicos, operativos y estructurales, entre los que se incluyen el tamaño de la empresa, el nivel de ventas, la inversión, los costos logísticos, la estructura productiva y la localización geográfica. Estas relaciones no suelen comportarse de manera lineal y, en muchos casos, presentan patrones difíciles de capturar mediante enfoques analíticos tradicionales; frente a esta realidad, la aplicación de técnicas de ciencia de datos y aprendizaje automático se justifica por su capacidad para modelar interacciones complejas y manejar distribuciones con alta dispersión, rasgos comunes en los montos exportados (Paredes, Vinces, Castro, & Llerena, 2025).

Desde el ámbito de la política pública, la importancia de esta investigación radica en su potencial para mejorar la focalización de los programas de promoción de exportaciones, de hecho, en la práctica, muchas iniciativas de apoyo se diseñan bajo esquemas generales que no siempre consideran las diferencias estructurales entre empresas, al contar con un modelo que estime la probabilidad de exportar y el volumen esperado de ventas externas, se dispone de una herramienta que permite priorizar intervenciones de manera más precisa, evidentemente, esta información puede orientar la asignación de recursos hacia empresas con mayor potencial de internacionalización y, al mismo tiempo, identificar grupos que requieren apoyos específicos para superar barreras vinculadas a inversión, tecnología o logística (Toaza & Rivera, 2025).

Debe señalarse que, la investigación se justifica además por su contribución al fortalecimiento de la toma de decisiones empresariales; para muchas empresas ecuatorianas, especialmente pequeñas y medianas, el proceso de internacionalización implica asumir riesgos significativos sin contar con información suficiente para evaluar sus posibilidades reales de éxito, es decir que, el modelo propuesto ofrece un insumo analítico que puede ser utilizado como referencia para evaluar la posición relativa de una empresa frente al mercado

externo, identificar debilidades internas y reconocer oportunidades de mejora, este aporte resulta relevante en un entorno donde la incertidumbre suele limitar la adopción de estrategias de expansión internacional (Pinos, Cevallos, Abril, & Pauta, 2025).

Otro elemento que refuerza la importancia del estudio es su contribución metodológica. La integración de un enfoque dual, que separa la estimación de la propensión exportadora del análisis del volumen exportado, responde a una distinción conceptual ampliamente reconocida en la literatura económica, esta separación permite abordar de manera más precisa dos decisiones que, aunque relacionadas, responden a dinámicas distintas, es por ello que, la aplicación de modelos de clasificación y regresión no lineal, junto con procedimientos de preprocesamiento adecuados, aporta una estructura metodológica robusta que puede ser replicada en otros contextos de análisis empresarial (Garzón y Diaz, 2023).

De manera complementaria, la justificación del estudio se fortalece también por el uso de información representativa del universo empresarial ecuatoriano, la incorporación de los pesos muestrales de la Encuesta Estructural Empresarial del INEC permite que las estimaciones obtenidas reflejen con mayor fidelidad la estructura real del tejido productivo, este aspecto resulta especialmente relevante cuando los resultados se utilizan para orientar decisiones de carácter institucional, ya que reduce el riesgo de sesgos derivados del análisis de muestras no ponderadas, la consideración explícita de la representatividad poblacional refuerza la validez de los hallazgos y amplía su utilidad práctica (Paredes, Vinces, Castro, & Llerena, 2025).

En el sentido del análisis, del ámbito académico, el estudio cobra relevancia al aportar al análisis sobre el uso de técnicas de aprendizaje automático en la investigación

económica; la propuesta se ubica en un espacio interdisciplinario donde convergen la teoría económica, la estadística y la ciencia de datos, lo que abre la posibilidad de examinar problemas clásicos del comercio internacional desde enfoques distintos, esta dinámica coyuntural, enmarca la utilidad para trabajos posteriores que busquen integrar fundamentos teóricos con herramientas predictivas orientadas a la toma de decisiones (Toaza y Rivera, 2025).

De esta manera, la relevancia del estudio también se manifiesta en su aporte al análisis territorial y sectorial del desempeño exportador, se plantea que, al trabajar con variables que capturan diferencias regionales y sectoriales, el modelo permite identificar patrones asociados a contextos productivos específicos, esta información resulta valiosa para el diseño de estrategias diferenciadas que consideren las particularidades de cada territorio y sector económico; de este modo, la investigación contribuye a una comprensión más matizada del proceso exportador y evita generalizaciones que pueden ocultar dinámicas relevantes a nivel local.

En el plano social y productivo, la justificación del estudio se relaciona con su potencial contribución al desarrollo económico, se propone identificar factores que favorecen la inserción internacional de las empresas, la investigación aporta insumos que pueden utilizarse para fortalecer el tejido productivo y promover una participación más amplia en el comercio exterior, por lo cual es evidente, que fortalecimiento tiene implicaciones positivas sobre la generación de empleo, la diversificación productiva y la estabilidad económica, especialmente en un país donde las exportaciones cumplen un papel central en la dinámica macroeconómica (Martínez y Cobos, 2025).

Por tal motivo, la importancia del estudio se refuerza además por su carácter aplicado y operativo; el modelo predictivo dual no se concibe únicamente como un ejercicio académico, sino como una herramienta que puede ser utilizada por distintos actores para apoyar procesos de análisis y planificación, es decir, que, la claridad en el diseño metodológico y la atención a la interpretabilidad de los resultados facilitan su adopción por parte de usuarios con distintos niveles de formación técnica, la perspectiva entonces, es que, la practicidad del modelo, amplía el impacto potencial de la investigación y favorece su utilización en contextos reales.

Por último, la justificación del estudio se apoya en su capacidad de adaptación y actualización. La metodología propuesta puede aplicarse a nuevas versiones de la encuesta utilizada o a otras fuentes de información empresarial, lo que permite mantener vigente el análisis a lo largo del tiempo, esta flexibilidad amplía el alcance del estudio y refuerza su valor como referencia metodológica para el análisis del comportamiento exportador, es así que, la investigación se posiciona como un aporte relevante tanto para el análisis académico como para la formulación de estrategias orientadas al fortalecimiento de la inserción internacional de las empresas ecuatorianas.

1.2.1 Impacto de la Investigación

El desarrollo de esta investigación genera un aporte significativo al análisis del comportamiento exportador de las empresas ecuatorianas, ya que propone una forma distinta de aproximarse a un fenómeno que tradicionalmente ha sido observado desde cifras agregadas. Al trabajar con información a nivel de firma, el modelo predictivo dual permite comprender con mayor detalle cómo interactúan las características económicas, productivas y estructurales en la decisión de exportar y en el volumen de ventas externas. Esta mirada

resulta especialmente pertinente en un país donde la estructura empresarial es diversa y donde conviven organizaciones con capacidades muy distintas, tanto en términos de escala como de acceso a recursos productivos y tecnológicos (Paredes, Vinces, Castro, & Llerena, 2025).

Desde la perspectiva de la gestión pública, el impacto del estudio se refleja en la posibilidad de mejorar la forma en que se diseñan y ejecutan las políticas de promoción de exportaciones; contar con estimaciones sobre la probabilidad de exportar y sobre el volumen esperado para cada empresa facilita una asignación más precisa de los recursos destinados a programas de apoyo; de este modo, las instituciones pueden dirigir sus esfuerzos hacia aquellas firmas que presentan condiciones favorables para ingresar a mercados internacionales, mientras se identifican con mayor claridad los grupos que enfrentan restricciones estructurales que requieren intervenciones diferenciadas; esta focalización contribuye a elevar la eficiencia del gasto público y a fortalecer la efectividad de las estrategias de internacionalización (Santa Cruz y Brito, 2023).

En el ámbito empresarial, el impacto se manifiesta a través de la generación de información que apoya la toma de decisiones estratégicas; es por esto que, las empresas pueden utilizar los resultados del modelo como una referencia para evaluar su posición frente al mercado externo, identificar brechas internas y reconocer áreas donde la inversión o la reorganización productiva podrían mejorar sus posibilidades de exportación, es evidente que el aporte adquiere especial relevancia para pequeñas y medianas empresas, que suelen operar con información limitada y enfrentan mayores niveles de incertidumbre al considerar procesos de internacionalización; disponer de un análisis basado en datos reales reduce ese margen de incertidumbre y aporta mayor claridad al proceso de planificación (Toaza y Rivera, 2025).

Por otra parte, la investigación tiene un impacto metodológico relevante al demostrar la utilidad de las técnicas de ciencia de datos y aprendizaje automático en el análisis económico aplicado; el uso de modelos de clasificación y regresión no lineal permite capturar relaciones complejas que difícilmente pueden representarse mediante enfoques econométricos tradicionales, sobre todo cuando se trabaja con distribuciones asimétricas y con alta presencia de valores extremos, como ocurre con los montos exportados, desde este enfoque se contribuye a ampliar el repertorio de herramientas disponibles para el análisis empresarial y ofrece una referencia práctica para investigaciones que enfrentan desafíos similares en términos de estructura de datos (Garzón y Diaz, 2023).

Cabe señalar que, la incorporación de los pesos muestrales provenientes de la Encuesta Estructural Empresarial del INEC refuerza el impacto del estudio al asegurar que las estimaciones obtenidas reflejen de manera más fiel la estructura real del universo empresarial ecuatoriano; este aspecto resulta fundamental cuando los resultados se utilizan para orientar decisiones de política pública, ya que evita interpretaciones sesgadas que podrían derivarse de análisis no ponderados, la consideración explícita de la representatividad poblacional fortalece la credibilidad del modelo y mejora su utilidad como insumo para la planificación y evaluación de programas orientados al fortalecimiento del sector externo (Investigación y Marketing, 2023).

Desde una mirada académica, el estudio contribuye a la discusión sobre la heterogeneidad empresarial y el comercio internacional al articular bases teóricas con herramientas actuales de análisis de datos, lo que facilita un mayor acercamiento entre la teoría económica y aplicaciones empíricas orientadas a la predicción, lo que crea un vínculo entre planteamientos conceptuales y requerimientos prácticos, además de abrir la posibilidad de desarrollar investigaciones posteriores que profundicen en el uso de modelos predictivos

aplicados a otras dimensiones del desempeño empresarial, ampliando así el campo de análisis más allá del ámbito exportador.

En la esfera productiva y social, el aporte de la investigación se relaciona con su capacidad para fortalecer el tejido empresarial; al reconocer patrones vinculados a la participación en mercados internacionales, el estudio ofrece insumos que pueden servir de base para el diseño de programas de capacitación, asistencia técnica y apoyo financiero orientados a empresas con potencial exportador, este reconocimiento favorece una inserción internacional más equilibrada y aporta a la reducción de brechas entre sectores y territorios, con efectos positivos sobre el desarrollo económico y la generación de empleo.

De igual manera, el modelo dual permite estimar el valor esperado de las exportaciones tanto a nivel individual como agregado, lo que amplía el impacto del estudio hacia la planificación estratégica de mediano plazo; las estimaciones poblacionales derivadas del modelo pueden utilizarse para analizar escenarios de crecimiento exportador y para evaluar el alcance potencial de distintas estrategias de promoción, sin embargo, la investigación no persigue objetivos causales, las predicciones generadas constituyen un insumo valioso para el análisis prospectivo y para la formulación de estrategias basadas en evidencia (Ordoñez, Pinos, & García, 2020).

Bajo este planteamiento, la investigación busca comprender y anticipar el comportamiento exportador de las empresas ecuatorianas desde una perspectiva empírica y orientada a la toma de decisiones; la atención puesta en la predicción, la representatividad de la información y la lectura interpretativa de los resultados refuerzan su utilidad para actores del sector público, privado y académico, además de consolidar su contribución al análisis del desarrollo productivo y a la inserción internacional del país.

1.3 Alcance de la Investigación

La investigación se orienta al estudio del comportamiento exportador de las empresas ecuatorianas mediante el desarrollo de un modelo predictivo dual que permite analizar, de forma diferenciada, la probabilidad de participación en mercados internacionales y el volumen potencial de exportaciones; por esta razón, el alcance se define a partir del uso de información empresarial proveniente de la Encuesta Estructural Empresarial del INEC, lo que posibilita trabajar con datos representativos del tejido productivo nacional y observar con mayor detalle las diferencias existentes entre empresas de distintos tamaños, sectores económicos y ubicaciones geográficas. Este enfoque permite centrar el análisis en la firma como unidad de estudio y comprender cómo sus características influyen en el desempeño exportador (Toaza y Rivera, 2025)

Dentro de este marco, el estudio abarca el análisis de variables económicas, operativas y estructurales disponibles en la base de datos seleccionada, entre ellas ventas, activos, personal ocupado, niveles de inversión, costos asociados al transporte, sector de actividad y localización territorial, por lo mismo, estas variables se utilizan como insumos para la construcción de los modelos predictivos, con el propósito de identificar patrones relevantes asociados tanto a la decisión de exportar como a la magnitud del monto exportado, por ello, el alcance del estudio se concentra en capturar relaciones complejas entre estas características empresariales, reconociendo que el desempeño exportador no responde a dinámicas simples ni uniformes, sino a interacciones múltiples que varían según el contexto productivo y regional.

En definitiva, la investigación comprende el diseño e implementación de un modelo dual basado en técnicas de ciencia de datos y aprendizaje automático, consecuentemente, en

una primera etapa, se desarrolla un modelo de clasificación orientado a estimar la propensión exportadora de las empresas, considerando la existencia de un desequilibrio entre firmas exportadoras y no exportadoras dentro del universo analizado, posteriormente, el estudio se enfoca en proyectar el monto exportado por aquellas empresas que ya participan en el comercio exterior, mediante el uso de modelos de regresión no lineal capaces de manejar distribuciones asimétricas y concentraciones de valores elevados. Ambas etapas forman parte integral del alcance metodológico y se articulan para generar una medida integrada del potencial exportador a nivel de empresa (Martínez y Cobos, 2025).

En consecuencia, el alcance del estudio incluye de manera explícita las tareas de preparación y tratamiento de los datos, tales como la limpieza de registros, la imputación de valores faltantes, la transformación de variables y el manejo de valores extremos, estas acciones se realizan con el objetivo de asegurar la calidad de la información utilizada y mejorar el desempeño de los modelos predictivos, en adelante, se incorpora el uso de los pesos muestrales proporcionados por el INEC, los cuales permiten que las estimaciones reflejen con mayor fidelidad la estructura real del universo empresarial ecuatoriano, la aplicación de estos pesos forma parte del alcance tanto en el proceso de entrenamiento como en la evaluación de los resultados (Ordoñez et al., 2020).

Dentro de este orden, la evaluación del desempeño de los modelos constituye otro componente central del alcance de la investigación; es así que, el estudio contempla el uso de métricas específicas para cada etapa del modelo dual, así como indicadores agregados que permiten valorar la precisión de las estimaciones a nivel individual y poblacional; este proceso incluye la validación de los modelos mediante esquemas de partición de los datos y el análisis comparativo entre los valores observados y las predicciones generadas. El alcance se extiende también al análisis del valor esperado de exportaciones, entendido como la

combinación entre la probabilidad estimada de exportar y el monto proyectado, lo que ofrece una medida integrada del potencial exportador de cada empresa (Paredes et al., 2025).

Estas consideraciones, la investigación se circunscribe al análisis predictivo y no persigue establecer relaciones de causalidad entre las variables examinadas; tampoco se orienta a identificar efectos directos de políticas públicas específicas ni a evaluar impactos derivados de intervenciones institucionales concretas, de la misma manera, el estudio no incorpora análisis longitudinales ni información de panel, por lo que los resultados se limitan al periodo cubierto por la base de datos empleada, por ende, estas delimitaciones definen con claridad el alcance del trabajo y permiten encuadrarlo dentro de un objetivo preciso y factible, propio de una investigación aplicada en el ámbito de la ciencia de datos.

Por consiguiente, tampoco se encuentra dentro del alcance la elaboración de proyecciones macroeconómicas de largo plazo ni la simulación de escenarios futuros a nivel agregado del país; el énfasis del estudio se sitúa en el nivel de la empresa, con el propósito de generar una herramienta que apoye procesos de análisis y decisión en contextos donde resulta necesario identificar potencial exportador a partir de información observable; en este sentido, el modelo desarrollado se concibe como un instrumento de apoyo a la planificación estratégica, tanto para entidades públicas como para organizaciones privadas interesadas en comprender mejor el comportamiento exportador del tejido empresarial (Aldrin et al., 2024).

Dentro de este marco, el alcance del estudio incluye además un componente interpretativo orientado a facilitar la comprensión de los resultados obtenidos, se consideran herramientas que permiten identificar la relevancia de las variables utilizadas y analizar cómo influyen en las estimaciones de probabilidad y monto exportado, cabe resaltar que, este análisis busca mejorar la utilidad práctica del modelo y favorecer su uso por parte de actores

que no necesariamente cuentan con formación técnica en aprendizaje automático, la interpretación se aborda desde una perspectiva aplicada, con énfasis en la lectura económica de los resultados y en sus posibles implicaciones para la gestión empresarial y la formulación de políticas.

En esta perspectiva, la investigación se concibe como un estudio replicable y adaptable a futuras aplicaciones; el alcance metodológico contempla la posibilidad de extender el modelo a nuevas versiones de la Encuesta Estructural Empresarial o a otras fuentes de información empresarial, siempre que se disponga de variables comparables, por lo que, esta característica amplía el valor del estudio y permite que la metodología propuesta sea utilizada como base para investigaciones posteriores o para ejercicios de actualización periódica del análisis exportador, manteniendo la coherencia con el enfoque predictivo adoptado.

En consecuencia, el alcance de la investigación se define, por tanto, por su orientación aplicada, su enfoque a nivel de empresa y su énfasis en el uso de técnicas de ciencia de datos para el análisis del comercio exterior, el estudio se limita a generar estimaciones y análisis que apoyen la comprensión del potencial exportador de las empresas ecuatorianas, manteniendo una lógica metodológica clara y transparente, sin extrapolar conclusiones más allá de los datos disponibles ni del periodo analizado.

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar un modelo predictivo dual que estime la propensión exportadora y el volumen esperado de exportaciones de las empresas ecuatorianas mediante técnicas de ciencia de datos aplicadas a la información empresarial del INEC.

1.4.2 Objetivos Específicos

Identificar las variables que influyen en la propensión exportadora mediante un modelo de clasificación basado en aprendizaje automático.

Estimar el monto exportado por las empresas participantes en comercio exterior mediante un modelo de regresión no lineal con preprocesamiento especializado.

Evaluar el desempeño del modelo dual mediante métricas individuales y agregadas para generar una herramienta útil en la priorización de decisiones empresariales y de política pública.

Capítulo II

2. Revisión de la Literatura

2.1 Estado del Arte

En los últimos años la manera de exportar de las empresas ha evolucionado de una forma significativa, siendo impulsada por el desarrollo e implementación de métodos de analítica avanzada conjuntamente trabajados con una gran cantidad de datos empresariales.

Diversas investigaciones actuales reconocen que las formas para internacionalizar las exportaciones empresariales no son lineales, sino que son una consecuencia de decisiones internas empresariales, o incluso factores del mercado actual dependiendo del área o producto de exportación.

Por tal motivo, la prospección exportadora hace referencia a la decisión que realiza una empresa para incursionar en mercados internacionales y el volumen exportado hace referencia a la magnitud de dicha incursión.

Esta doble evaluación facilita el entendimiento de dos componentes importantes en el mundo exportador internacional, el margen extensivo y el margen intensivo del comercio.

La internacionalidad de empresas comprende un elemento clave en el desarrollo de países llamados emergentes, como es el caso de Ecuador, en el cual el acceso a mercados internacionales es vital para la diversificación de la producción, y mejorar de manera significativa la competitividad.

Además, adaptarse a las nuevas dinámicas de exportación permite a los encargados de tomar decisiones y académicos diseñar políticas focalizadas en el incremento y sostenibilidad de empresas ecuatorianas (BCE, 2020).

2.1.1 Teoría de la Heterogeneidad de Firmas y Selección de la Exportación.

De acuerdo a Bernard, Jensen (BEJK) Meltiz a inicios de los años dos mil, los conceptos modernos del comercio internacional admiten la heterogeneidad de empresas, este concepto fue pionero en su tiempo, donde propone que las empresas se diferencian en aspectos importantes como productividad, estructura de capital y estrategias en el mercado, las mismas que generan la autoselección, demostrando que solo las empresas más grandes, fuertes y con más productividad pueden convertirse en empresas exportadoras, ya que estas lograrían cubrir los altos costos que comprende el mercado internacional, las mismas que deben cubrir un umbral de costos fijos de entrada y costos variables presentes en el comercio internacional, tomando en consideración que este tipo de empresas deben contar con certificaciones, logística, capacidad tecnológica, seguros, requisitos regulatorios, etc. mientras que las empresas más pequeñas o menos productivas estarían destinadas al mercado local (Bernard et al, 2012, pp. 1-25).

Esta teoría se ha reafirmado realizando estudios con empresas estadounidenses que demuestran que no solo tienen la capacidad de exportación, sino que también pagan salarios más altos a sus empleados, este patrón se ha implementado en múltiples países incluido Ecuador.

2.1.2 Determinantes Observables de la Participación y de la Intensidad Exportadora

Cada país presenta ciertos factores determinantes observables, sin embargo, internacionalmente hay consensos en los estudios de comercio exterior.

2.1.2.1 Tamaño de la Empresa

Las empresas más grandes, generalmente medidas por el número de empleados, productividad o ventas, sobrepasan el umbral de costos unitarios, tienen capacidad de minimizar riesgos y cuentan con infraestructuras especializadas como logística o comercio exterior, que facilitan poner su nombre en el panorama internacional.

En Ecuador esto se hace muy notorio al comparar empresas grandes con empresas que se manejan en el ámbito local como aquellas de tipo A y B (Melitz, 2003).

2.1.2.2 Productividad

Debido a que en algunos casos no se cuenta con indicadores de productividad, es común emplear medidas alternativas como las ventas por empleado, el valor agregado generado por cada trabajador, la eficiencia operativa o los márgenes de utilidad.

Algunos estudios señalan que la productividad, medida a través de estos proxies, conforma el principal factor que determina el comportamiento exportador de las empresas. Las empresas que muestran mayores niveles de productividad son capaces de asumir los elevados costos fijos requeridos para adaptarse y competir en mercados internacionales (Wagner, 2007).

2.1.2.3 Capital, Inversión y Tecnología.

La posibilidad de invertir, disponer de suficiente capital y adoptar nuevas tecnologías fortalece la competitividad de las empresas. En áreas como la manufactura, la minería o los servicios especializados, contar con maquinaria actualizada, certificaciones internacionales, herramientas digitales y procesos de automatización, influye de manera directa el poder acceder a mercados externos y sostener actividades de exportación (Helpman, 2008).

2.1.2.4 Costos de Transporte y Localización Geográfica.

Los gastos asociados a la logística y al transporte afectan de manera significativa en la rentabilidad de las exportaciones. Aquellas empresas ubicadas lejos de puertos o de centros estratégicos de distribución, generalmente enfrentan mayores costos adicionales, lo que disminuye sus posibilidades de exportar. En el caso de Ecuador, las diferencias regionales entre la Sierra, la Costa y la Amazonía se traducen en variaciones en infraestructura, nivel de acceso y costos logísticos (BCE, 2020).

2.1.2.5 Sector Económico (CIIU).

El sector al que pertenece una empresa influye de forma significativa en su capacidad para exportar. Por ejemplo, los sectores que requieren de más inversión en maquinaria y equipos tienen una demanda internacional alta o producen bienes únicos con estándares especiales que suelen exportar más. Gracias a la clasificación CIIU utilizada en la base de datos de este trabajo, se puede ver claramente que sectores como la manufactura, la minería, la construcción, el comercio y los servicios muestran comportamientos exportadores muy distintos entre sí (Helpman, 2008).

2.1.3 Distinción Clave: Participación (Binaria) vs. Monto (Intensidad Continua)

Según la literatura más reciente, la decisión de exportar y el volumen que se exporta son dos cosas diferentes, aunque estén relacionadas. La participación en el comercio internacional es una elección puntual, es decir, la empresa decide si exporta o no, y eso depende de la comparación que haga entre sus costos (tanto fijos como variables) y productividad. Por eso, normalmente se utiliza un tipo de modelo estadístico pensado para decisiones que solo tienen dos opciones (Álvarez, 2005, pp. 589–607)).

Por otro lado, el monto exportado es una variable continua que se analiza sólo después de que la empresa haya decidido exportar. No tiene sentido calcular esta variable para empresas que no exportan, así que solamente se observa en las que sí participan en el comercio internacional. El volumen depende de cosas como cuántos mercados diferentes atiende la empresa, cuántos productos exporta, las estrategias comerciales que usa, si tiene financiamiento, su capacidad instalada y qué tan eficiente es su logística.

Separar estos dos aspectos es importante para evitar errores en el análisis y para construir modelos predictivos que sean más exactos. Por esto, la investigación utiliza modelos de “dos etapas”, donde cada parte se estudia por separado (López, 2009).

2.1.4 Modelos Tradicionales Relevantes

2.1.4.1 Modelos de Participación Exportadora (Logit y Probit)

Para analizar la participación exportadora de las empresas, los modelos Logit y Probit son los que más se han usado tradicionalmente porque están diseñados para decisiones que tienen solo dos respuestas posibles, como exportar o no exportar. Estos modelos calculan la probabilidad de que una empresa exporte tomando en cuenta características propias como su tamaño, productividad, inversión, nivel de capital humano y aspectos específicos del sector o la región (Bernard, 1999).

La mayor ventaja de estos modelos es que son fáciles de interpretar, ya que permiten ver cómo cambia la probabilidad de exportar si una variable aumenta y también pueden mostrar relaciones de riesgos relativos. Esto es útil para hacer recomendaciones o diagnósticos de políticas públicas. Sin embargo, tienen varios problemas: suponen que la relación entre las variables y la decisión de exportar es lineal en la escala log-odds, son muy sensibles si algunos predictores se parecen mucho entre sí, no reflejan bien las interacciones ni las relaciones no lineales, además, necesitan que los errores sigan ciertas reglas estadísticas bastante estrictas.

Por eso, cuando lo más importante es lograr buenas predicciones (como ocurre en los modelos predictivos duales), los algoritmos de machine learning que son más flexibles suelen tener mejores resultados que los modelos tradicionales (Wooldridge, 2010).

2.1.4.2 Modelos de Monto: Tobit y Heckman.

Para analizar cuánto exportan las empresas, normalmente se usan regresiones lineales aplicadas al logaritmo del monto exportado. Esta transformación ayuda a que los datos sean menos desiguales, controla mejor la variabilidad y hace que los resultados sean más fáciles de

interpretar; por ejemplo, podemos entender cómo un cambio en una variable afecta porcentualmente el monto exportado.

Sin embargo, la regresión lineal tiene varios problemas, sobre todo cuando la base de datos incluye valores extremos, muchos datos concentrados en pocos casos, o cuando las relaciones entre las variables no son lineales. Esto pasa mucho en el comercio internacional, donde hay pocas empresas que exportan muchísimo y otras que exportan muy poco o nada (Heckman, 1979, pp. 589–607).

Por lo tanto, en la literatura también se han usado modelos Tobit y Heckman:

i. El modelo Tobit sirve cuando la variable volumen está limitada en cero, y supone que todas las empresas podrían exportar, pero algunas no lo hacen por restricciones. El problema es que, en realidad, los ceros en las exportaciones suelen significar que la empresa sencillamente no exporta, no que estén censurados, lo que puede generar errores al interpretarlo.
ii. El modelo Heckman reconoce que solo tenemos datos de montos exportados para las empresas que sí exportan. Primero estima la probabilidad de exportar y después corrige el posible sesgo cuando analizamos el monto que exportan, ayudando a controlar diferencias ocultas y problemas de interpretación.

Estos modelos comparten varios inconvenientes para el objetivo de predecir: las relaciones son lineales, captan poco las interacciones complicadas, son muy sensibles a valores extremos, piden que los datos cumplan reglas estadísticas bastante exigentes (como normalidad y homocedasticidad) que en la práctica suelen fallar (Heckman, 1979).

2.1.4.3 Modelos Two-Stage / Hurdle

En los últimos años, los modelos de dos etapas, también conocidos como hurdle models, se han vuelto bastante comunes en el estudio del comercio internacional a nivel de

empresas. Estos modelos separan el análisis en dos partes: por un lado, miden la probabilidad de que una empresa empiece a exportar (participación) y, por otro, analizan cuánto exporta una vez ha tomado la decisión de participar.

Este enfoque se basa muy bien en la teoría de empresas heterogéneas, que explica por qué no todas las empresas exportan y por qué existen diferencias entre quienes sí lo hacen. En la práctica, trabajar con modelos de dos etapas permite obtener resultados más precisos y evitar errores cuando se usa solo una regresión lineal simple (OLS).

Para los estudios aplicados y la elaboración de políticas de exportación, este tipo de modelos ofrece información mucho más detallada porque ayuda a detectar qué factores hacen posible que las empresas accedan al mercado internacional, y qué aspectos influyen luego en su desempeño exportador. Esto es especialmente útil en el contexto ecuatoriano, donde la cantidad de empresas exportadoras es pequeña en relación con el total, y donde unas pocas compañías concentran la mayor parte del volumen exportado (Cragg, 1971).

2.2 Marco Teórico

2.2.1 Modelos de Clasificación

En los últimos años, los métodos de machine learning han mostrado mejores resultados para predecir qué empresas van a exportar. Entre los modelos que más se usan están:

i. Random Forest, que es muy resistente al ruido, no necesita asumir que los datos siguen ciertas reglas estadísticas y puede detectar relaciones y variaciones no lineales entre variables (Breiman, 2001).
ii. Gradient Boosting y XGBoost, que suelen ser los que dan mejores resultados en las predicciones porque se enfocan en ir corrigiendo los errores y son capaces de

encontrar patrones complejos (Chen 2016, pp. 785–794).

iii. Support Vector Machines (SVM), que funcionan muy bien cuando tenemos muchos datos y variables, especialmente si se usan técnicas como los kernels para captar relaciones no lineales (Cortes & Vapnik, 1995).

Estos modelos de machine learning funcionan mejor que los métodos tradicionales porque son mucho más flexibles y pueden captar relaciones complicadas entre variables, lo que es útil en empresas de distintos sectores con características muy diversas, como ocurre con la clasificación CIIU en los estudios de exportación.

2.2.1.1 Modelos de Regresión no Lineal (RandomForestRegressor, XGBoostRegressor)

Para estimar cuánto exporta una empresa, los modelos de regresión que no son lineales tienen varias ventajas importantes. Por ejemplo, pueden identificar relaciones complicadas entre las variables sin que uno tenga que definir formas matemáticas específicas desde el inicio. Además, estos modelos no necesitan que los datos sean normales ni que tengan varianza constante, así que son menos exigentes con las características de la base de datos. Otra ventaja es que son bastante resilientes frente a valores extremos o casos fuera de lo común, y permiten analizar qué factores tienen más peso en las predicciones usando métodos como la importancia de Gini o los valores SHAP (Hastie et al, 2009).

Esto resulta muy útil en Ecuador, donde las diferencias entre sectores, tamaños de empresas y regiones son muy marcadas. Así, estos algoritmos logran capturar efectos combinados y relaciones que un modelo lineal tradicional no podría mostrar de manera adecuada.

2.2.1.2 Transformaciones y Manejo de Distribuciones Sesgadas

En los datos de empresas, es común encontrar variables como ventas, activos y montos exportados con distribuciones muy desiguales y valores extremos bastante altos. Para

hacerlos más manejables, se pueden aplicar varias técnicas: una de ellas es la transformación log1p (log(1+x)), que ayuda a trabajar mejor con datos que tienen ceros; otra es la winsorización, que reduce el efecto de los valores extremos sin perder datos; y también se usa la normalización o estandarización, muy útil cuando los modelos son sensibles a la escala de las variables (Huber, 1981).

Estas transformaciones hacen que los modelos sean más estables desde el punto de vista numérico y ayudan a mejorar la precisión de las predicciones.

2.2.1.3 Validación, Calibración y Evaluación

Para saber si un modelo predictivo funciona bien, es importante usar métodos de evaluación fiables. Por ejemplo, dividir los datos en entrenamiento y prueba (train/test) ayuda a ver cómo se desempeña el modelo con datos nuevos. La validación cruzada, como k-fold, sirve para comprobar la variabilidad de los resultados y reducir el riesgo de que el modelo se ajuste demasiado a la muestra (overfitting) (Kuhn, 2013).

Hay diferentes métricas para evaluar los modelos, como AUC-ROC, F1-score, precisión, recall y matriz de confusión para los modelos de clasificación; y RMSE, MAE y R^2 para los de regresión. Además, calibrar el modelo con curvas de confiabilidad es útil cuando queremos saber qué tan acertados son los valores de probabilidad que predice en clasificación.

En conjuntos de datos donde hay muchas más empresas que no exportan que exportadoras, técnicas como el re-muestreo y la regularización se vuelven esenciales para que el análisis sea válido. Usando todas estas herramientas, se puede comparar de forma objetiva los modelos tradicionales con los de ciencia de datos y así elegir el que mejor se adapta a cada parte: la predicción de participación y la predicción del monto exportado . (Varma, 2006).

2.2.2 Uso de Pesos Muestrales (f_exp)

2.2.2.1 Representatividad Poblacional

Cuando se hacen estudios usando datos de empresas, como encuestas o registros administrativos, casi nunca se tiene información de todas las empresas existentes, es decir, no suele ser un censo total. Por eso, se usan pesos muestrales para ajustar la muestra y hacer que represente mejor a la población real. En este contexto, el f_exp es el factor de expansión: un número que multiplica cada empresa incluida en el estudio para estimar cuántas empresas reales representa en el país.

El uso de f_exp tiene dos propósitos: por un lado permite hacer estimaciones correctas a nivel de toda la población. Si no usamos pesos muestrales, podríamos caer en errores y sesgos al calcular la proporción de empresas exportadoras por sector o el volumen total exportado, porque algunas empresas tienen más probabilidades de aparecer en la muestra debido a su tamaño, sector o región. El f_exp corrige estas diferencias y nos ayuda a obtener estimaciones más representativas.

Y, cuando los resultados del estudio se usan para apoyar políticas públicas (como programas para incentivar exportaciones o capacitaciones), es fundamental que los modelos reflejen lo que ocurre en la población real. Si no usamos pesos y simplemente sacamos promedios, podríamos terminar sobrevalorando o subestimando la importancia de ciertos grupos, por ejemplo, de microempresas o empresas que están fuera de las principales ciudades.

El factor de expansión es clave para poder trasladar los resultados que obtenemos de la muestra a todo el universo de empresas ecuatorianas. Esto es esencial si queremos ir más allá de saber qué empresas exportan dentro del estudio y queremos estimar realmente los efectos o los recursos que se necesitan en todo el país (Cochran, 1977).

2.2.2.2 Integrar Sample_Weight en Fit y en Cálculo de Métricas y Agregados Poblacionales

Cuando se entrenan modelos de clasificación o regresión, casi todas las librerías modernas permiten usar pesos en el ajuste del modelo. Se incluye sample_weight=f_exp quiere decir que cada observación en el entrenamiento se multiplica por su factor de expansión, lo que lleva a dos efectos clave: los datos con mayor peso tienen más influencia al calcular la función objetivo como el log-loss para clasificación o el MSE para regresión y los parámetros del modelo se ajustan para minimizar el error ponderado, logrando así predicciones que representan mejor a toda la población y no solo a la muestra escogida.

En scikit-learn, la mayoría de los modelos, como RandomForestClassifier o LogisticRegression, aceptan el argumento sample_weight al entrenar usando .fit(X, y, sample_weight=xxx).En XGBoost y LightGBM se puede pasar un vector de pesos cuando se crea el DMatrix o el Dataset, y también hay opciones de peso en la API para cada fila. En métodos bayesianos o M-estimators, muchos paquetes permiten agregar los pesos en la función de verosimilitud o en la función de pérdida.

Por eso, lo más recomendable es usar sample_weight=f_exp cuando se entrene el modelo, para que la predicción final sea realmente representativa de la población y no solo de la muestra. (Hastie et al, 2009)

2.2.2.3 Riesgos: Pesos Extremos y Mitigaciones

Cuando los pesos de las observaciones son muy diferentes entre sí, eso puede hacer que la muestra realmente útil se reduzca, porque unas pocas observaciones con mucho peso pueden influir exageradamente en los resultados y en cómo se ajusta el modelo.

En modelos muy flexibles, como los de boosting, existe el riesgo de que el modelo se adapte demasiado a estos casos con mucho peso, lo que podría empeorar su desempeño cuando se lo prueba con datos nuevos.

También puede haber problemas de inestabilidad numérica si los pesos son demasiado grandes o pequeños, y eso puede afectar los cálculos del modelo.

Por último, si una observación fuera de lo común (un outlier) tiene además un peso alto, terminará teniendo un efecto desmesurado tanto en las predicciones como en las estimaciones a nivel de toda la población (Little, 2019).

2.2.3 Interpretabilidad y Relevancia

2.2.3.1 Herramientas de Interpretabilidad

Es fundamental poder interpretar bien los modelos predictivos, sobre todo si sus resultados se van a utilizar para tomar decisiones de política pública. Es importante ofrecer interpretaciones tanto a nivel general (o sea, saber qué variables son las más importantes en promedio) como a nivel específico (entender por qué el modelo hace una predicción concreta para determinada empresa). Algunas herramientas y métodos útiles para esto son:

Importancia de Variables en Árboles de Decisión: Es una métrica que viene por defecto en modelos como Random Forest y Gradient Boosting, basada en cuánta "impureza" reduce cada variable. Sirve para tener una visión general, pero hay que tener cuidado porque puede estar sesgada si hay variables con muchas categorías o diferentes escalas.

Permutación Importante: Consiste en ver cuánto empeora el rendimiento del modelo si mezclamos aleatoriamente una variable. Es más general (compatible con cualquier modelo) y menos sesgada que la importancia por ganancia, así que suele ofrecer una medida más confiable de qué tan relevante es cada variable para las predicciones.

SHAP (SHapley Additive exPlanations): Este enfoque viene de la teoría de juegos y descompone cada predicción en el aporte de cada variable usando los valores SHAP. Tiene varias ventajas: explica de forma consistente y específica cada predicción individual, es eficiente de computar en modelos tipo árbol gracias a TreeSHAP, permite visualizar la dependencia, el resumen y las interacciones de las variables en los resultados.

Gráficos de Dependencia Parcial (PDP): Los gráficos de dependencia parcial (PDP) muestran cómo cambia en promedio la predicción del modelo a medida que una variable varía, manteniendo el resto de las variables constantes. Son útiles para ver relaciones que no sean lineales. Los gráficos ALE (Accumulated Local Effects) sirven como alternativa cuando las variables están correlacionadas, ya que solucionan el problema que tienen los PDP en esos casos (Lundberg & Lee, 2017).

2.2.3.2 Transparencia Para Decisiones Públicas, Priorización y Fairness

Los resultados de los modelos no deberían verse solo como un objetivo académico, sobre todo cuando se usan para tomar decisiones de política pública. En esos casos, que sean interpretables es súper importante para que las decisiones sean legítimas, transparentes y efectivas.

Algunas razones importantes:

i. Transparencia y responsabilidad: Las instituciones públicas tienen que poder explicar por qué eligen a ciertos beneficiarios para sus programas, como subsidios para exportar. Mostrar qué variables influyen en la selección ayuda a que el proceso sea menos opaco y mejore la confianza de todos.
ii. Mejor asignación de recursos: Si se entiende bien el score de propensión a exportar, los apoyos (como capacitaciones o créditos) se pueden dirigir a las empresas que más lo necesitan o a las que tengan más potencial, optimizando el

impacto.

iii. Equidad: Si no se revisan los resultados del modelo por grupos (por ejemplo, según tamaño de empresa, región o sector), se corre el riesgo de que el modelo termine reforzando desigualdades, como favorecer siempre a las empresas grandes o de ciudades. Interpretar bien el modelo permite detectar y corregir estos posibles sesgos.

iv. Diseño más efectivo de políticas: Al saber exactamente qué variables influyen (como acceso a crédito o inversión en tecnología), se pueden diseñar políticas complementarias que sean más útiles y dirigidas.

v. Comunicación: Métodos como SHAP o PDP facilitan mucho que tanto los responsables de las políticas como las propias empresas entiendan cómo funciona el modelo y por qué se hacen ciertas recomendaciones (Lundberg & Lee, 2017).

Como buenas prácticas se recomienda: mostrar cuáles variables son más importantes (a nivel general y con ejemplos concretos), analizar si el modelo es justo con diferentes grupos, sugerir acciones que puedan aumentar las posibilidades de exportar de una empresa (por ejemplo, a través de escenarios hipotéticos),y documentar todas las limitaciones o suposiciones del análisis (Barocas et al, 2019).

2.2.4 Elección Metodológica

Estrategia que utiliza dos modelos distintos:

Un clasificador para predecir la probabilidad de que una empresa sea exportadora.

Un modelo de regresión para estimar el monto exportado, pero solo entre las empresas que ya exportan.

A esto le sumamos transformaciones de las variables, como aplicar log1p a los montos para trabajar con ceros para evitar que las empresas que exportan montos extremadamente altos influyan demasiado en los resultados.

Durante el entrenamiento y cálculo de estimaciones para la población total, usaremos sample_weight = f_exp, que es el factor de expansión para que cada empresa cuente de acuerdo con su peso poblacional. Además de los siguientes criterios:

i. Teoría y evidencia: Separar el margen de decisión (quién exporta) del margen de intensidad (cuánto exporta) sigue la teoría de Melitz y es lo que la mayor parte de los estudios aplicados recomienda (modelos de dos etapas). Por eso, la estrategia dual es sólida y está bien respaldada conceptualmente.
ii. Metas prácticas: El objetivo es poder predecir a nivel individual qué empresas tienen más posibilidades de exportar (útil para políticas públicas) y también saber cuánto se podría exportar en total o cuántas empresas estarían involucradas. El uso de f_exp se asegura de que esos resultados realmente representen la realidad nacional.
iii. Resistencia a valores extremos: Al transformar los montos, evitamos que un puñado de empresas gigantescas desestabilicen la predicción, logrando así mejores resultados y más estabilidad en casos nuevos.
iv. Predicción más precisa y explicable: La literatura muestra que los modelos de aprendizaje automático como RandomForest o XGBoost suelen predecir mejor que los modelos clásicos, y además soportan el uso de sample_weight y métricas como feature_importances y SHAP, lo que ayuda a interpretar las decisiones del modelo.
v. Reducción de riesgos: Finalmente, normalizar, revisar los pesos y hacer pruebas de sensibilidad ayuda a evitar inestabilidad numérica y otros problemas que podrían surgir por el diseño de la muestra.

Capítulo III

3. Desarrollo del Trabajo

3.1 Enfoque General del Estudio

En el presente proyecto se desarrolla un modelo dual que se compone por:

i. Un modelo de clasificación que tiene como objetivo determinar si una empresa ecuatoriana es exportadora o no exportadora de acuerdo con las características económicas y productivas que también se desarrollarán en este capítulo
ii. Un modelo de regresión que tiene como objetivo predecir el valor anual de exportación, en caso de que la empresa sea clasificada como exportadora.

Se ha utilizado este enfoque debido al alto desequilibrio que presenta el sector exportador a nivel nacional y a la limitación de la predicción del valor exportado ya que solo tiene sentido en los casos en que las unidades de negocio efectivamente realicen exportaciones. Este modelo permite capturas ambas dinámicas, manteniendo precisión estadística y coherencia entre estos dos componentes predictivos.

El análisis toma datos de la Encuesta Estructural Empresarial ENESEM 2023, publicada por el Instituto Ecuatoriano de Estadística y Censos (INEC). La encuesta recopila datos relevantes de empresas ecuatorianas en los ámbitos de producción, empleo, capital, utilidades, bienes intermedios, formación bruta de capital fijo, entre otros.

3.1.1 Justificación del Modelo Dual

En la variable ‘Valor de Exportación (v1005)’ se encontraron las siguientes características estructurales que complican su análisis directo:

i. El valor de exportación en la mayoría de las empresas es cero, es decir que no exportan
ii. La distribución de la variable es fuertemente asimétrica y presenta valores extremos.

Por lo cual se optó por utilizar un enfoque dual, con el fin de tratar este problema de manera consistente con la literatura económica y con casos similares de economías altamente heterogéneas, la predicción se divide en dos etapas:

a) Predicción de la probabilidad de exportación $P(exporta)$

b) Predicción del monto de exportación condicionado a las empresas que exportan $E(valor|exporta)$ utilizando transformación logarítmica y winsorización para mitigar el impacto de valores extremos.

El resultado $\left(\hat{y}_{esperado}\right)$ Se obtiene de la multiplicación de la probabilidad de exportar con el monto esperado de exportación.

$$\hat{y}_{esperado} = P(exporta) \times E(\widehat{monto|exporta})$$

3.2 Datos Utilizados

3.2.1 Fuente de Datos y Muestra

Se utilizó la Encuesta Estructural Empresarial ENESEM 2023, esta base de datos recopila información de empresas y datos sobre su producción, insumos, empleo, ventas, entre otros. En total la base tiene más 4.400 registros de empresas, con el procesamiento se identificaron que del total aproximadamente un 23% de empresas son exportadoras es decir que el valor de exportación es mayor a 0.

3.2.2 Variable Objetivo Para el Clasificador

La variable v1005 en la base de datos es el monto de exportación por empresa, a partir de esta se generó una variable binaria:

$$exportador = \{1 \quad si\ v1005 > 0\ 0 \quad si\ v1005 = 0$$

3.2.3 Variable Objetivo Para la Regresión

Para utilizar la variable con el objetivo de predecir el monto exportado por empresa se utilizó la siguiente transformación:

$$y = log\ (1 + monto\_winsorizado)$$

De esta manera

i. Los valores de la variable 'monto exportado' fueron winzorizados, es decir se realizó un ajuste de los valores extremos al percentil $p = 0.99$ para mitigar este impacto en el modelo

ii. Se empleó el algoritmo natural (log) al valor ajustado y, se sumó uno para los casos en que el valor transformado sea cero o cercano a cero.

Es importante aclarar que la winsorización se llevó a cabo únicamente con los datos de entrenamiento. Para garantizar que la evaluación del modelo sea frente a datos de prueba no modificados y de esta manera evitar fugas de información.

3.2.4 Variables Predictoras

Para predecir el valor de exportación por empresa se seleccionaron variables económicas, productivas, geográficas y estructurales relevantes y con potencial para predecir la probabilidad y el monto de exportación.

Tabla 1.
Categorías de las Variables Predictoras

Categoría	Detalle de las variables
Localización y estructura empresarial	● *Provincia* (ubicación geográfica) ● *cod_tamano, des_tamano* (tamaño empresarial) ● *cod_sector, des_sector* (sector productivo empresarial) ● *anio_ruc_dis* (antigüedad aproximada)
Variables financieras y productivas	● Ventas: *v1001, v1002, v1003, v1004* ● Comercializables (producto interno bruto empresarial): *Vbp, Vbc, Vns* ● Activos fijos: *v4001* ● Remuneraciones y personal: *totalpeoc, totremun*

Variables operativas y de servicios	● Inversión fija y adquisiciones: *adqvnv, totadquisi, fbk, fbkf_1* ● Consumo de energía y agua: *cant_ener, cant_agua* ● Transporte y logística: *v1125, v1126*
Ponderador muestral	En el análisis este factor se utiliza como peso muestral (sample_weight) con el fin de corregir sesgos asociados a diseño muestral y asegurar representatividad de los resultados a nivel poblacional.

Nota: La descripción detallada de todas las variables incluidas en esta tabla se encuentra disponible en el Anexo A

3.3 Análisis Exploratorio de Datos (EDA)

El análisis estará enfocado en los siguientes aspectos clave:

i. La distribución de la variable monto de exportación (v1005)

ii. Identificar valores atípicos en las variables

iii. Analizar la proporción de empresas exportadoras de acuerdo con su sector, provincia y tamaño

iv. Identificar correlaciones fundamentales para la construcción del modelo predictivo

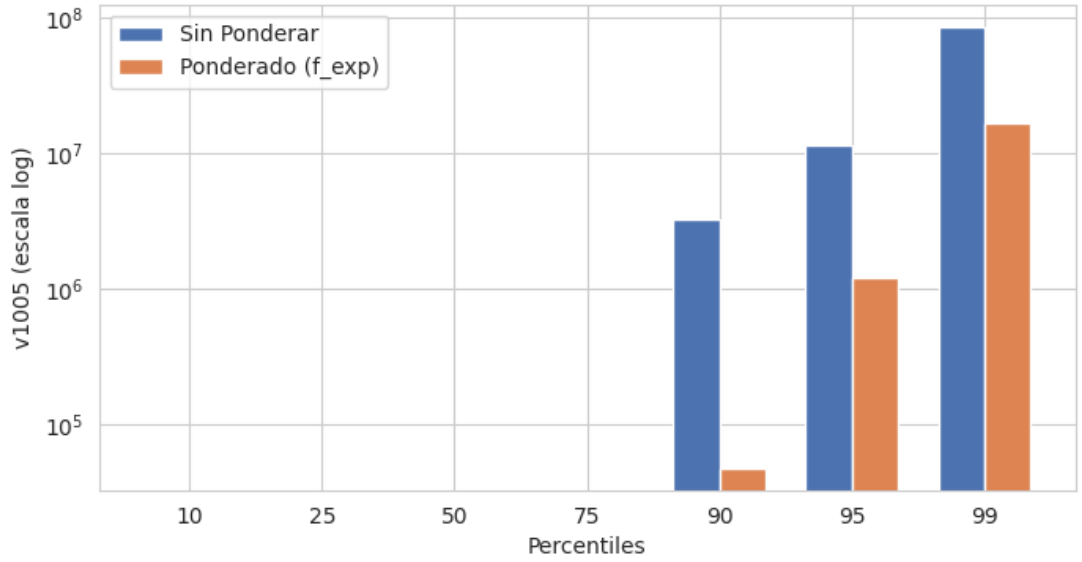
v. Generar diagnósticos específicamente relevantes para el pipeline de predicción.

3.3.1 Estadísticas Descriptivas y Percentiles del Monto Exportado

Se elaboró un análisis detallado de percentiles ($p10\ hasta\ p99$) del monto exportado ($v1005$) con el propósito de comprender su distribución, tanto de la muestra original y también aplicando el factor de expansión (f_exp) para estimaciones poblacionales, en la figura se evidencia una comparativa visual de estos conjuntos de datos en escala logarítmica.

Figura 1.

Percentiles v1005: Muestra y Población (Ponderado)



Nota: Elaboración propia basada en la ENESEM (INEC, 2023)

En la muestra de datos no ponderados se puede visualizar que, los montos de exportación tienen una alta concentración en los percentiles más altos. En el caso del $p99$ este valor supera los 85 millones USD, y los valores máximos incluso llegan a los cientos de millones, es decir hay una cola de datos muy marcada en la distribución, estos valores extremos evidencia que pocas empresas concentran la mayor parte de exportaciones.

En la muestra de datos con valores ponderados con el factor de expansión (f_exp) los percentiles indican valores mucho menores, retomando el caso del $p99$, el monto descendió a 16.6 millones USD, por consiguiente se concluye que, las empresas que presentan valores de exportación extremadamente altos tienen un bajo peso muestral, y la estructura de la población empresarial está representada por unidades de negocio con monto de exportación más moderados.

3.3.2 Análisis de la Tasa de Exportación por Segmentos Estructurales

Se analizaron la proporción de empresas exportadoras usando los datos originales y también los datos ponderados con el factor de expansión para muestras poblacionales.

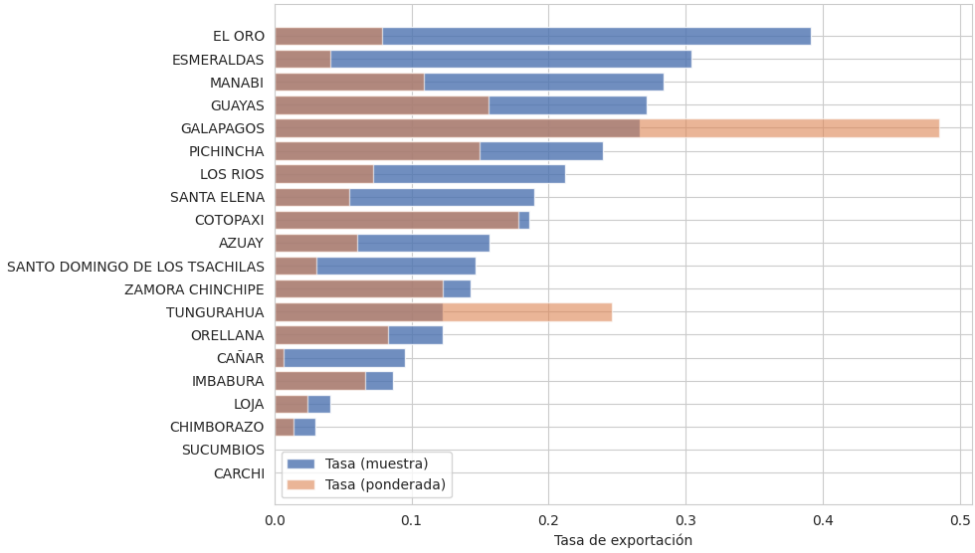
3.3.2.1 Tasa de Exportación Global

Del total de registros la tasa de empresas que registran exportaciones es del 23.53%. No obstante, al ponderar los registros con el factor de expansión esta proporción desciende a 12.22%, es decir en términos relativos los datos sin ponderar sobre estiman la representación de las empresas exportadoras.

3.3.2.2 Tasa de Exportación por Provincias

Como se observa en la gráfica las provincias con mayores tasas de exportación con registros no ponderados son El Oro (39%), Guayas (27%) y Pichincha (24%). Sin embargo al aplicar la ponderación con el factor de expansión otras provincias toman un mayor peso como en el caso de Galápagos, ya que en este caso aunque son pocos registros estos tienen un fuerte peso muestral.

Figura 2.
Tasa de Exportación por Provincias



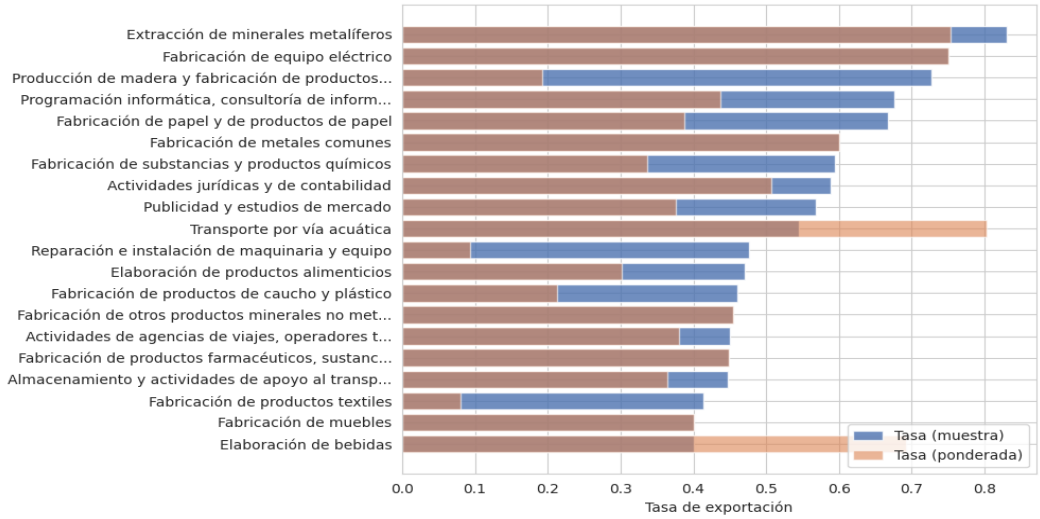
Nota: Elaboración propia basada en la ENESEM (INEC, 2023)

3.3.2.3 Tasa de Exportación por Sector

La tasa de exportación cambia considerablemente entre los distintos sectores productivos, entre los sectores que presentan las mayores tasas de empresas exportadoras se encuentran las empresa de la Producción de Madera y Derivados, la Reparación e Instalación de Maquinaria y Equipo, y la Fabricación de Productos Textiles, mientras que en el sector de Fabricación de Otros Productos minerales casi no se registran empresas exportadoras, es decir el sector productivo es una variable determinante para este análisis.

Figura 3.

Tasa de Exportación por Sector



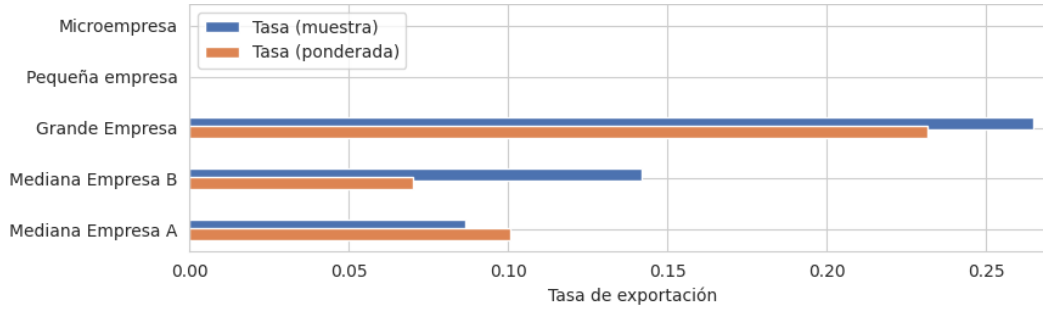
Nota: Elaboración propia basada en la ENESEM (INEC, 2023)

3.3.2.4 Tasa de Exportación por Tamaño de Empresa

En la gráfica se visualiza que, las empresas grandes e intermedias tienen una mayor proporción de empresas exportadoras, esto generalmente se relaciona con su mayor capacidad productiva, infraestructura o mayores niveles de ventas, mientras que en pequeñas y microempresas el registro de exportaciones es casi nulo. Por esta marcada diferencia entre los diferentes tamaños empresas se concluye que también es una variable fundamental en el análisis de exportación.

Figura 4.

Tasa de Exportación por Tamaño de Empresa



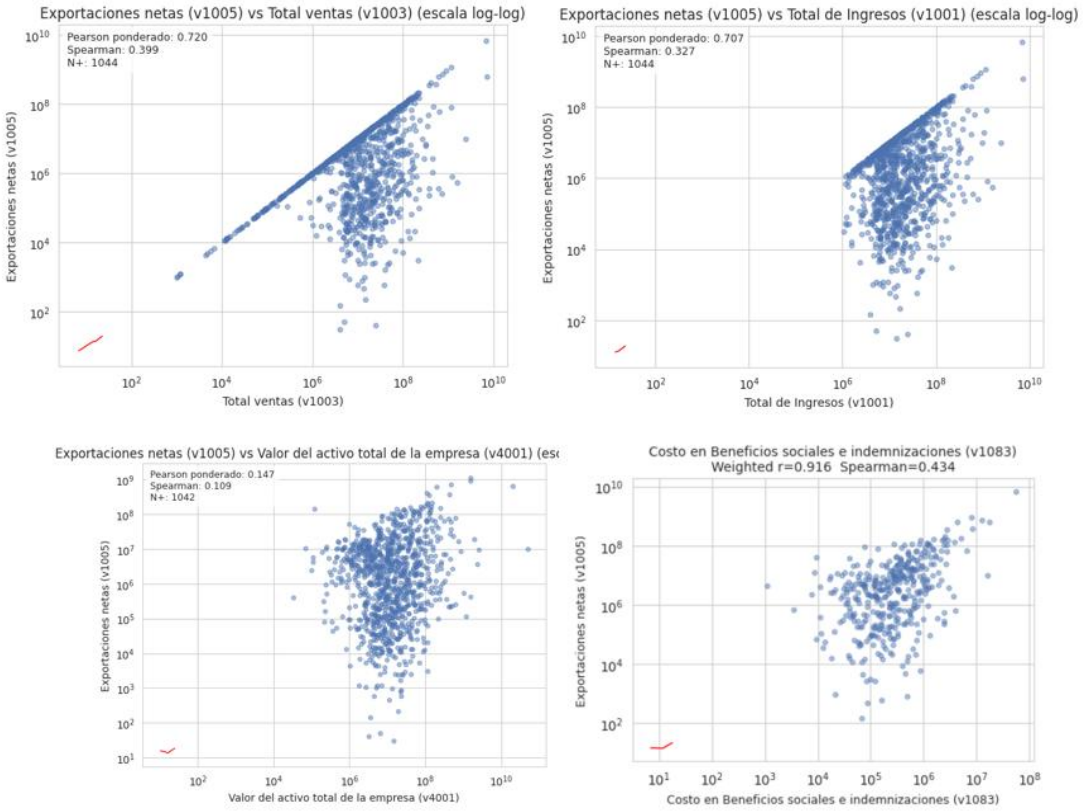
Nota: Elaboración propia basada en la ENESEM (INEC, 2023)

3.3.3 Análisis de Correlación de las Variables

En este análisis se relacionó la variable objetivo 'monto de exportación' con las variables cuantitativas del modelo, siendo estas principalmente de ventas, costos y producción. Además se presentan indicadores de correlación de Pearson para su evaluación y posterior interpretación. En las variables que presentan una relación más relevante las gráficas se presentan en escala log-log con el fin de facilitar su interpretación visual.

Figura 5.

Correlación de las Variables Predictoras con la Variable Objetivo



Nota: Elaboración propia basada en la ENESEM (INEC, 2023)

Como se evidencia en la matriz de correlación las principales variables relacionadas al monto de exportación son variables de producción y ventas como el valor agregado empresarial ($r = 0.949$), los ingresos, las ventas, costos por beneficios sociales, costos por salarios y gastos asociados a transportes, la producción total empresarial ($r = 0.71$).

Por el contrario variables como valor del activo total presentan un coeficiente de Pearson muy bajo ($r = 0.109$), lo que indica una relación débil y poca influencia en la variable de exportaciones.

Figura 6.
Matriz de Correlaciones

Nota: Elaboración propia basada en la ENESEM (INEC, 2023)

En resumen las variables que cuantifican la productividad y la capacidad económica de las empresas son las que presentan una mayor correlación con la variable de exportaciones, las variables relacionadas con costos logísticos como el transporte también tienen un potencial explicativo frente a la variable objetivo, mientras que variables de estructura y empleo indican una relación más moderado pero positiva y las variables de pasivos e inversiones muestran poca relación con las exportaciones presentando coeficientes de correlación menores a 0,10.

3.3.4 Preparación y Limpieza de los Datos

En este apartado se explica el tratamiento que se dio a los datos con el fin de garantizar la calidad de la información y disminuir sesgos que puedan afectar la calidad del modelo de predicción.

3.3.4.1 Estandarización de Datos Monetarios

Se implementó la función ($clean\_numeric\_series$) con el objetivo de estandarizar las columnas que contenían información referente a datos monetarios como por ejemplo ventas, exportaciones e ingresos ya que en algunos el símbolo de dólar variaba y se presentaban caracteres no numéricos. En este pasó también se unificó el separador de decimales y la función también trata de forma segura como valores faltantes a las celdas vacías.

3.3.4.2 Clasificación de Variables Cuantitativas y Cualitativas

En el modelo se identifica explícitamente las variables cod_tamano, cod_ciud, cod_sector y $provincia$ ya que a pesar de tener datos numéricos estos corresponden a una codificación de variables categóricas, que podrían confundirse y en caso de ser tratadas como numéricas generarían correlaciones o llevarían a interpretaciones equivocadas.

3.3.4.3 Imputación de Valores Faltantes

Se aplicaron diferentes técnicas de imputación para los diferentes tipos de variables:

Variables cuantitativas: Se usó el valor de la mediana para datos numéricos faltantes, la ventaja de esta técnica frente al uso de la media es que es más robusta en casos de valores extremos y presenta menos afectación en distribuciones fuertemente asimétricas.

Variables cualitativas: Se añadió una nueva categoría “missing” de esta manera se evitó sesgar datos faltantes hacia la moda y también captar patrones que se relacionen con la falta de información.

3.3.4.4 Escalamiento de Datos Numéricos

Se utilizó StandarScaler en todas las variables numéricas para centrar y escalar estas variables. A pesar de que el modelo Random Forest presenta mayor robustez frente a los

datos sin escalamiento a diferencia de otras técnicas de regresión lineal, el escalamiento de los datos evita que variables de gran escala controlen de forma inadecuada el proceso de optimización.

3.3.4.5 Codificación de Variables Cualitativas Mediante One-Hot Encoding

Se utilizó la codificación One-Hot-Encondindg para generar cada categoría de las variables cualitativas a columnas binarias, esta codificación es estandarizada y evita que se asigne un orden específico a variables categóricas que no lo tienen. Además en este proceso se implementó la opción handle_unknown="ignore" para que en caso de que se genere una nueva categoría que no estaba en el entrenamiento no falle el modelo.

3.3.4.6 Split Estratificado por Exportador

La división de la base de datos se realizó mediante un split estratificado tomando como criterio de estratificación a la variable $exportador$, donde $exportador = 1$ si $v1005\_exportaciones > 0$. Este proceso es fundamental para mantener la misma proporción de exportadores en el conjunto de datos de entrenamiento y de prueba. Además se estableció que el tamaño del conjunto de datos de prueba corresponde al 20% y una semilla fija de 42.

Proceso de estratificación

- Tamaño total de la muestra: N = 4 436 observaciones.
- División aplicada: 80 % entrenamiento / 20 % prueba.

Resultados obtenidos:

- Conjunto de entrenamiento (TRAIN):

  n_train = 3 548 observaciones (≈ 80.0 % del total).

- Conjunto de prueba (TEST):

n_test = 888 observaciones (≈ 20.0 % del total).

Estratificación por exportador:

- Exportadores en TRAIN:

  n_export_train = 835, proporción ≈ 23.5344 %.

- Exportadores en TEST:

  n_export_test = 209, proporción ≈ 23.5360 %.

En este sentido se logró mantener la proporción de exportadores en los dos conjuntos de datos, garantizando la comparabilidad para evitar sesgos en las métricas de evaluación. Además el volumen de datos de entrenamiento (835 observaciones) es representativa para la construcción del modelo de regresión y el uso de la semilla fija asegura una adecuada reproducibilidad.

3.4 Pipeline de Clasificación (P(exporta|X))

Dentro del desarrollo del modelo dual, la primera etapa inicia consiste en estimar la probabilidad de que una unidad económica (empresa) registre exportaciones. Bajo esta lógica, la variable objetivo se planteó a través de una variable binaria:

$$exportador_i = \{1;\ si\ v1005\ > 0\ 0; si\ v1005 = 0$$

La variable binaria define 1 si el establecimiento declara un valor positivo en la variable v1005 (monto de exportaciones) y 0 en el caso de que no. Cabe destacar que, dicho planteamiento responde a que el proceso de exportar maneja características de comportamiento desbalanceadas, no lineales, y dependiente de interacciones entre múltiples variables, por esto, la justificación de generar un modelo de clasificación.

3.4.1 Estimador: Clasificador de Bosque Aleatorio (Parámetros Base).

Para el desarrollo del modelo dual se inició con la etapa de clasificación, donde se seleccionó como estimador a Bosque Aleatorio puesto que esta técnica de aprendizaje automático reduce el riesgo de sobreajuste y funciona con variables categóricas eficazmente, lo cual, asegura un desempeño sólido y fiable. A su vez, representa una técnica de aprendizaje por conjuntos (ensemble), lo que significa que combina diversos árboles de decisión generando una predicción final más robusta. El modelo incluye los parámetros de 200 árboles, una profundidad máxima de 16 niveles.

3.4.2 Entrenamiento Ponderado Mediante el Factor de Expansión

La data empleada para el presente estudio tiene datos provenientes de encuestas en base a un muestreo por lo que el factor de expansión $f_x$constituye un aspecto fundamental en el proceso del modelado debido a que, este peso representa la probabilidad de inclusión de cada unidad muestral (empresas) para estimar al universo poblacional.

Por tanto, al utilizar el factor de expansión a través del parámetro *sample_weight* en el entrenamiento se logra que el modelo garantice su probabilidad de inclusión y representatividad poblacional, lo cual generó que el modelo optimizará su proceso de aprendizaje ponderando para cada observación según su representatividad.

3.4.3 Validación del Modelo y Calibración

El método seleccionado para evaluar el desempeño del clasificador consiste en una validación cruzada estratificada (pliegues) que determina que cada subconjunto utilizado en el entrenamiento y prueba posee la misma proporción de empresas definidas tanto como exportadoras como no exportadoras que el conjunto total. Por tanto, se empleó esta validación cruzada (*StratifiedKFold*) para asegurar que cada fold (subconjuntos de entrenamiento y prueba) mantenga la proporción real de exportadores y se eviten sobrestimaciones.

En base a que los modelos de clasificación fundamentados en árboles como lo es el Bosque Aleatorio, por lo general, generan probabilidades de pertenencia a una clase que en ocasiones pueden estar ligeramente sesgada, por lo que, se consideró la implementación de una calibración opcional para que reflejen una precisión en la probabilidad real del evento. En consecuencia se evaluó el uso de *CalibratedClassifierCV* en base al método de regresión isotónica puesto que corresponde a una técnica de calibración no paramétrica que ajusten las probabilidades de salida de un clasificador.

3.5 Pipeline de Regresión (E[monto | exporta, X])

El componente de regresión dentro del modelo dual se empleó con el objetivo de estimar el monto exportado por parte de aquellas empresas que ya han superado la barrera de participación en mercados internacionales. La literatura económica reconoce que la distribución del valor exportado maneja niveles altos de asimetría, puesto que, la mayoría de las empresas exportan montos reducidos y un pequeño número aquellas concentran niveles excepcionalmente altos de exportación.

Esta dinámica, documentada en su estudio Heterogeneidad del Modo de Exportación y Productividad Empresarial: Una Prueba sobre el Aprendizaje al Exportar por (Bernard, 2011), también se observa de forma clara en la Encuesta Estructural Empresarial, donde los percentiles superiores del monto exportado (v1005) muestran saltos abruptos en los valores, especialmente a partir del percentil 95. Debido a esta estructura, la modelación del monto no puede realizarse directamente sobre la variable en escala original, puesto que, produciría un modelo altamente inestable, dominado por valores extremos.

3.5.1 Estimador: Regresor de Bosque Aleatorio

Para modelar el monto exportado, se implementó un regresor de bosque aleatorio, configurado con 300 árboles. Este modelo fue seleccionado por su estabilidad, su bajo riesgo

de sobreajuste y su capacidad para incluir interacciones no lineales entre las características productivas de las empresas.

Cabe destacar, la inclusión del factor de expansión como *sample_weight* durante el entrenamiento permite que los exportadores de mayor representatividad en la estructura poblacional concentren una mayor influencia en la función objetivo. Este paso es de suma relevancia, puesto que, sin la consideración de los pesos, el modelo revelaría patrones de la muestra, mas no de la población total de empresas del país.

3.5.2 Validación y Métricas

La evaluación del modelo se realizó únicamente sobre exportadores presentes en el conjunto de prueba. Los resultados típicos obtenidos por modelos de esta naturaleza suelen reflejar lo siguiente:

Tabla 2.
Métricas de desempeño del modelo de regresión

Métrica	Valor observado	Interpretación
MAE	USD 5,317,424	El error absoluto medio indica que, en promedio, el modelo presenta una desviación de aproximadamente **5.3 millones de USD** al estimar el monto exportado por una empresa exportadora individual. Este valor refleja la elevada dispersión existente en los montos de exportación.
RMSE	USD 43,785,499	El error cuadrático medio es considerablemente mayor que el MAE, lo que evidencia la fuerte penalización asociada a errores en empresas con volúmenes de exportación muy elevados, confirmando la presencia de una distribución con cola superior de la distribución.
R^2 (escala original)	0.375	El modelo logra explicar cerca del 37.5 % de la variabilidad real del monto exportado entre las empresas exportadoras, un resultado consistente con problemas económicos de alta complejidad estructural.
R^2 (escala logarítmica)	0.700	En la escala logarítmica, el poder explicativo del modelo se incrementa de forma significativa, alcanzando aproximadamente el

		70 %, lo que valida el uso de la transformación logarítmica del objetivo durante el entrenamiento.
MAE ponderado (f_exp)	≈ USD 6.0 millones	Al ponderar por representatividad poblacional, el error promedio aumenta, indicando que las empresas de mayor peso económico presentan una mayor dificultad de predicción.
RMSE ponderado (f_exp)	≈ USD 45.0 millones	El incremento del RMSE ponderado confirma que los mayores errores se concentran en un número reducido de empresas de gran escala, que dominan el valor total exportado del país.

Estos valores presentados son consecuentes con la presencia de elevada heterogeneidad por parte del sector exportador ecuatoriano y con la dificultad de predecir montos exactos en presencia de empresas extremadamente grandes. En este tipo de problemas, valores de R^2 en el rango de 0.30 y 0.50 son considerados adecuados y figuran una capacidad sólida del modelo para evidenciar patrones relevantes.

Cabe destacar que, el análisis de residuos mostró que la mayor parte del error se concentra en empresas de gran escala, mientras que para la mayoría de los exportadores medianos y pequeños, las predicciones generadas son bastante estables y cercanas a los valores reales.

3.5.3 Guardado del Modelo y del Preprocesador

Finalmente, el modelo de regresión y el preprocesador asociado fueron estructurados empleando la librería *Joblib*. Esta práctica es fundamental dentro del ciclo de vida de la ciencia de datos, puesto que permite preservar el estado completo del modelo y las diferentes transformaciones aplicadas, lo cual asegura su trazabilidad metodológica.

La serialización del modelo hace posible el desarrollo eficiente del modelo para su uso posterior en diversos escenarios: la inferencia en nuevas encuestas empresariales y mantener su actualización anual, la implementación en sistemas destinados para programas

de política pública, y la simplificación para el análisis de monitoreo y posterior impacto del tejido empresarial exportador. Por tanto, mantener una estructura previamente guardada permite reproducir el pipeline de la predicción sin necesidad de repetir el proceso de entrenamiento del modelo.

3.6 Combinación y Predicción Final

La etapa final del modelo dual integra las dos partes centrales del enfoque: la probabilidad estimada de que una empresa sea categorizada como exportadora y el monto esperado condicionado a que en efecto exporte. Este proceso combinatorio permite generar un valor esperado de exportación para cada empresa, lo cual resulta sumamente útil para la aplicación de política pública y para generar una estimación del potencial exportador del país. En el contexto ecuatoriano, caracterizado por una fuerte heterogeneidad productiva este enfoque resulta metodológicamente sólido y empíricamente pertinente.

3.6.1 Fórmula Operacional del Valor Esperado

Una vez entrenados y validados ambos componentes del modelo —el clasificador y el regresor— la predicción final para cada empresa se obtiene aplicando una formulación que combina ambas partes:

$$\hat{P}_i = classifier.predict_{proba(X_i)[1]}$$

$$\hat{E}_i = exp\big(regressor.predict(X_i)\big) - 1$$

$$\hat{Y}_i = \hat{P}_i \cdot \hat{E}_i$$

donde:

- $\hat{P}_i$es la probabilidad predicha de que la empresa $i$sea exportadora,
- $\hat{E}_i$es el monto esperado de exportación condicional a que exporte,

- $\hat{Y}_i$representa el valor esperado total de exportación para la empresa i, incluso si su probabilidad de exportar es baja o nula.

3.6.2 Evaluación del Valor Estimado

La evaluación del valor esperado se realizó comparando $\hat{Y}_i$ con el valor real reportado en la encuesta en la variable (*v1005*), donde se emplearon las métricas correspondientes de error absoluto medio (MAE) y error cuadrático medio (RMSE) las cuales fueron calculadas sobre todo el conjunto de prueba (incluidos los no exportadores). Esta evaluación conjunta es más exigente que las evaluaciones por separado de los modelos, ya que incorpora:

- La capacidad del clasificador para identificar correctamente como exportadores.
- La capacidad del regresor para aproximar el monto exportado.
- La coherencia en conjunto de la predicción integrada.

Cabe destacar que, la estimación del total esperado de exportaciones se generó a través del valor esperado de cada empresa ponderado por su factor de expansión:

$$Total\ Estimado = \sum_{i}^{n} \hat{Y}_i \cdot f\_exp_i$$

Por tanto, este indicador permite comparar la proyección del potencial exportador del modelo frente al total reportado por la encuesta, contribuyendo a la validación externa del mismo.

Tabla 3.
Métricas de evaluación del valor estimado

Métrica	Valor observado	Interpretación
MAE (no ponderado)	USD 1,356,120	En promedio, el modelo presenta un error absoluto de aproximadamente 1.36 millones de USD al predecir el valor exportado por empresa, considerando exportadores y no exportadores.

RMSE (no ponderado)	USD 21,457,964	El valor elevado del RMSE refleja la fuerte penalización asociada a errores en empresas con volúmenes de exportación muy altos, evidenciando la asimetría del fenómeno.
MAE ponderado (f_exp)	USD 420,775	Al ponderar por representatividad poblacional, el error promedio disminuye, lo que indica un buen ajuste del modelo en términos agregados y una adecuada captación del patrón poblacional.

Los resultados mostrados en la tabla muestran que el modelo dual reproduce con alta precisión el volumen exportador agregado, con una discrepancia inferior al 5%, lo cual, es destacable considerando la complejidad estructural del sector exportador y la presencia de empresas extremadamente grandes las mismas que generan una gran parte del volumen total. En términos generales, el modelo destaca un buen desempeño para la mayoría de empresas, concentrando sus errores más relevantes en la cola superior de exportaciones (grandes grupos).

3.6.2.1 Notas prácticas Sobre Winsorization y Prevención de Fuga de Datos

Un aspecto fundamental para la correcta ejecución del modelo dual corresponde a la forma en que se aplica la winsorización y medidas para evitar cualquier forma de filtración de información (*data leakage*) entre los conjuntos de entrenamiento y prueba. Cabe señalar que, este tipo de filtración, por lo general, ocurre cuando información correspondiente al conjunto de prueba influye, directa o indirectamente, en las transformaciones durante la etapa de entrenamiento, lo que generaría un modelo metodológicamente inválido.

En el caso de la Encuesta Estructural Empresarial (ENESEM), el monto exportado mantiene valores extremadamente altos (atípicos) en la cola superior de la distribución. Con la finalidad de controlar el impacto de estos valores atípicos (sin eliminarlos), se optó por aplicar una winsorización del 95%, que limita los valores mayores al percentil 95 a un límite calculado. No obstante, este umbral debe calcularse únicamente con los datos del conjunto de

entrenamiento, específicamente con los exportadores, debido a que, incluir datos del conjunto de prueba permitiría que información futura influya en las decisiones del modelo.

Bajo esta lógica, el límite superior de winsorización se obtuvo exclusivamente a partir de las empresas exportadoras que formaron parte del conjunto de entrenamiento. Una vez estimado este umbral, se aplicó uniformemente a los datos del conjunto de prueba sin recalcularlo.

De igual forma, la transformación logarítmica aplicada posterior a la winsorización: se determina y valida únicamente con data del entrenamiento, y posterior se aplica de manera consistente sobre los datos de prueba. Cabe mencionar que, mantener esta separación es indispensable para modelos no lineales como Random Forest o XGBoost, puesto que estos algoritmos son sensibles a cambios y pueden capturar patrones erróneos si se ven expuestos a información externa al entrenamiento.

En conjunto, la correcta secuencia desde la winsorización para el entrenamiento, posterior transformación logarítmica y separación absoluta entre entrenamiento y prueba asegura la transparencia, validez y rigor metodológico del modelo dual aplicado a datos empresariales. En otras palabras, esta estructuración sistemática evita sesgos, reduce sobreajuste y permite que las métricas obtenidas reflejen de manera precisa la capacidad predictiva del modelo.

3.6.3 Hiperparametrización y Validación Cruzada Ambos Modelos

La etapa de hiperparametrización y validación cruzada constituye uno de los componentes metodológicos más importantes para el desarrollo del modelo dual. Puesto que, su propósito es garantizar que los algoritmos seleccionados (tanto para la clasificación como para la regresión) no solo ajusten adecuadamente los datos de entrenamiento, sino que mantengan un comportamiento robusto y estable frente a nuevos datos.

Este proceso se vuelve imprescindible cuando se trabaja con data de ámbito empresarial como la Encuesta Estructural Empresarial (ENESEM), la cual, por lo general presenta variabilidad estructural, alta heterogeneidad productiva, presencia de valores atípicos y un desequilibrio entre empresas exportadoras y no exportadoras. Por tanto, la validación cruzada permite evaluar distintos conjuntos de hiperparámetros, reduciendo el sobreajuste y mejorando la capacidad predictiva.

3.6.4 Parámetros y Rangos Evaluados

En primer lugar, para ambos modelos (clasificador y regresor) se definió un conjunto de hiperparámetros que influyen directamente en la capacidad de aprendizaje del algoritmo. Los parámetros fueron seleccionados siguiendo referencias de la literatura como evidencia empírica obtenida mediante pruebas.

Entre los principales hiperparámetros considerados se incluyeron el número de árboles (n_estimators), la profundidad máxima (max_depth), el número mínimo de observaciones para considerar una división (min_samples_split), el número mínimo de muestras por hoja (min_samples_leaf), y la cantidad de variables evaluadas en cada división (max_features).

En el caso específico de los modelos basados en XGBoost, se incorporaron hiperparámetros adicionales asociados a la dinámica del boosting y a la regularización. Entre ellos se incluyó la tasa de aprendizaje (*learning_rate*), que controla la magnitud de la actualización en cada iteración y permite un aprendizaje más gradual; el parámetro de submuestreo (*subsampl*e), que reduce la dependencia entre árboles consecutivos; y la fracción de variables utilizadas por árbol (*colsample_bytree*), que introduce variabilidad adicional. Finalmente, se consideró el parámetro gamma que actúa como un mecanismo explícito de regularización estructura

Tabla 4.
Parámetro Evaluados

Parámetro	Valores evaluados	Modelos aplicados
n_estimators	100, 200, 300, 500	Random Forest (Clasificador y Regresor)
max_depth	8, 12, 16, None	Random Forest / XGBoost
min_samples_split	2, 5, 10	Random Forest
min_samples_leaf	1, 2, 4	Random Forest
max_features	“auto”, “sqrt”, 0.5	Random Forest
learning_rate	0.01, 0.05, 0.1	XGBoost
subsample	0.7, 0.8, 1.0	XGBoost
colsample_bytree	0.6, 0.8, 1.0	XGBoost
gamma	0, 1, 5	XGBoost

La revisión de múltiples combinaciones de estos parámetros permitió evaluar la sensibilidad de los modelos y seleccionar la configuración que maximiza el desempeño en las métricas definidas para cada etapa: el ROC_AUC en clasificación y el R^2, MAE y RMSE en regresión. Cabe mencionar que, se mantuvo el uso del mismo preprocesador en todas las combinaciones evaluadas, lo cual asegura que las comparaciones entre configuraciones sean consistentes.

3.6.4.1 Validación Cruzada Estratificada Para Clasificación y KFold Para Regresión

Un paso clave para el diseño del modelo dual corresponde la implementación de esquemas adecuados de validación cruzada para cada tipo de modelo, las cuales consisten:

i. Para la etapa de clasificación: se utilizó StratifiedKFold, la cual garantiza que cada partición conserve la misma proporción de empresas exportadoras y no exportadoras que el conjunto original. Tomando en consideración, la necesidad de mantener la estructura del desequilibrio de clases, puesto que, una distribución desbalanceada en los pliegues podría devenir en modelos inestables y métricas no representativas. En el

contexto de la Encuesta Estructural Empresarial (ENESEM), donde los exportadores representan alrededor del 20% de las empresas, su uso resultó indispensable.

ii. Para la etapa de regresión, al tratarse únicamente de empresas exportadoras, la validación cruzada se realizó mediante KFold sin estratificación. Esto se debe a que todas las observaciones dentro del conjunto ya cumplen la condición categórica de ser exportadoras y, por lo tanto, la estratificación no aporta beneficios adicionales.

3.6.4.2 Manejo de Pesos Muestrales en la Búsqueda de Hiperparámetros

Una particularidad relevante del presente estudio de investigación es que todos los modelos fueron entrenados empleando los pesos muestrales (f_exp), los cuales determinan la probabilidad de inclusión de cada empresa en la muestra de la encuesta.

Para la incorporación de estos pesos dentro del proceso de validación cruzada y la búsqueda de hiperparámetros, se utilizaron los argumentos *fit_params* dentro de *GridSearchCV* y *RandomizedSearchCV*. Por tanto, la inclusión de los pesos muestrales en cada pliegue (folds) de la validación cruzada garantiza que la optimización del modelo esté alineada con la estructura del muestreo.

Esto implica que los errores cometidos por el modelo se retribuyen de manera proporcional a cada empresa dentro del universo empresarial del país. Cabe destacar, ignorar los pesos generaría un sesgo estructural, en otras palabras, las empresas con baja probabilidad de inclusión, pero alto peso poblacional mantendría un impacto insuficiente en la función de costo del modelo, lo cual, generaría predicciones no representativas.

En síntesis, la hiperparametrización definida por validación cruzada estratificada, junto con el uso correcto de pesos muestrales (factor de expansión), permitió seleccionar configuraciones robustas para ambos modelos. Este proceso garantiza que el modelo final

tenga un desempeño predictivo correcto, y mantenga representatividad poblacional, reproducibilidad y coherencia estadística.

3.6.5 Interpretabilidad y Diagnóstico del Modelo Dual

Tras la obtención de los modelos finales de clasificación y regresión es indispensable evaluar su interpretabilidad y comportamiento interno. En el contexto de este estudio, la interpretabilidad permitió identificar qué variables determinan la probabilidad de exportar, cuáles impulsan el monto exportado y qué tipos de empresas concentran los principales errores del modelo. Por tanto, este diagnóstico es clave para comprender la estructura productiva nacional.

3.6.5.1 Importancias Globales de Variables en el Clasificador y el Regresor

La primera herramienta utilizada para interpretación del modelo dual fue el análisis de importancia global de variables, en base, a las medidas proporcionadas por los algoritmos Random Forest y XGBoost. Este grado de importancia se calcula en función de la reducción promedio de pérdida (error) que genera cada variable al ser empleada en divisiones de los árboles durante el entrenamiento.

Para el clasificador, las variables con mayor importancia global coincidieron con las revisadas previamente en la literatura económica sobre la participación exportadora, entre ellas destacaron:

i. Ventas totales (*v1001–v1004),* utilizadas como una medida indirecta de productividad y capacidad de generación de ingresos, las cuales incrementan la probabilidad de inserción en mercados internacionales.

ii. Tamaño de la empresa *(cod_tamano),* donde las empresas medianas y grandes presentan una mayor probabilidad de participar en actividades exportadoras debido a economías de escala y mayor capacidad organizativa.

iii. Sector económico *(cod_ciiu2d),* destacándose actividades manufactureras y sectores con orientación histórica hacia el comercio exterior.

iv. Remuneraciones totales *(totremun)* y empleo total *(totalpeoc),* variables que reflejan la escala productiva y el nivel de operación de la empresa, estrechamente asociados a la probabilidad de exportación.

v. Inversión fija *(fbk, adqvnv),* relacionada con la capacidad de sostener procesos productivos complejos y operaciones internacionales de forma continua.

Estas variables coinciden con los factores encontrados en estudios como (Orellana, 2025) (Bernard, 2011) e investigaciones en América Latina que muestran que la productividad, el tamaño y la intensidad de capital son los principales predictores de la participación exportadora.

Para el regresor, que predice el monto exportado entre exportadores, destacaron como variables más influyentes:

i. Ventas brutas (Vbp)

ii. Activos y capital productivo (v4001, fbk)

iii. Variables de insumos logísticos y energéticos (v1125, v1126)

iv. Inversión fija reciente (fbkf_1)

v. Sector económico como descriptor de la estructura productiva.

En esta etapa, las variables de insumos logísticos y de inversión cobran mayor relevancia que en la clasificación, lo que coincide con la lógica económica: una mayor capacidad de exportación depende intensamente de la capacidad de producción e infraestructura.

3.6.5.2 Interpretabilidad por Permutación (Permutation Importance) con Scoring Ponderado

Para complementar la interpretación y evitar cualquier tipo de sesgo se aplicó Permutation Importance, el cual corresponde a un método robusto que mide el aumento del error cuando el valor de una variable se altera de manera aleatoria. Este análisis se realizó tanto para el clasificador como para el regresor, utilizando métricas ponderadas por el factor de expansión f_exp, lo cual garantiza que la importancia refleje la estructura poblacional.

Los resultados esperados para el clasificador constituyen incrementos en el error al permutar variables como tamaño, ventas y sector, lo que conlleva a su efecto determinante. En contraposición, para el regresor las variables relacionadas con escala productiva (producción, activos, energía) muestran los mayores incrementos de error, validando su efecto en la capacidad de exportación de la empresa.

3.6.5.3 Explicabilidad con SHAP: Análisis Global e Individual

Para garantizar una explicabilidad a mayor detalle del modelo, se utilizó la metodología SHAP (Shapley Additive exPlanations), la cual, es considerada actualmente como el estándar más robusto para interpretar modelos complejos (modelo dual). Este método consiste en la asignación para cada variable una contribución cuantitativa a cada predicción, lo que permite analizar tanto tendencias generales como casos individuales.

A nivel global, los valores SHAP confirmaron los patrones detectados en: el tamaño, las ventas y el sector, los cuales fueron los principales impulsores positivos de la probabilidad

de exportar, mientras que la escasez de inversión reciente redujo esta probabilidad para empresas pequeñas. En el caso del monto exportado, las variables productivas y logísticas mostraron contribuciones altas, especialmente en empresas con niveles elevados de producción y capital fijo.

A nivel individual, SHAP permitió analizar empresas específicas y comprender el por qué algunas tuvieron predicciones altas o bajas. Por ejemplo, empresas que tengan valores elevados de inversión en maquinaria, alta dotación energética y elevado nivel de ventas tuvieron barras SHAP positivas que incrementaban significativamente el monto esperado. En contraposición, empresas pequeñas con niveles bajos de inversión y actividad logística mostraron contribuciones negativas que redujeron el monto esperado.

Este nivel de explicabilidad es particularmente valioso para las empresas, puesto que permite justificar la implementación en programas de desarrollo, evitando decisiones opacas con el fin de incrementar su capacidad de exportación.

3.6.5.4 Análisis de Residuales del Regresor y Patrones en la Cola

El análisis de residuales del regresor evalúa la diferencia entre el monto real exportado y la predicción estimada con el fin de evaluar la estabilidad del modelo y comprender en cuáles segmentos se concentran mayoritariamente los errores. Por tanto, al examinar los residuales, se observó que el modelo tiene un buen desempeño para la mayoría de las empresas exportadoras, manteniendo errores relativamente pequeños entre niveles bajos y medios de su capacidad de exportación.

Sin embargo, tal como se revisó en la literatura sobre comercio exterior, los principales errores se concentran en la cola superior de la distribución, es decir, en las empresas que exportan montos excepcionalmente altos. Estas empresas, aunque pocas en

número, representan una porción significativa del comercio total del país, y origina que su comportamiento sea altamente difícil de modelar debido a su heterogeneidad y tamaño.

Los residuales en la cola muestran una dispersión elevada, lo que se refleja en métricas como el RMSE (regularmente sensible a valores extremos). Esto no representa un fallo del modelo, sino una característica marcada en el contexto económico donde: los grandes exportadores son extremadamente diversos y están influenciados por factores estratégicos, financieros y de mercado que por lo general no siempre se reflejan en los datos recolectados en encuestas estructurales.

Aun así, el modelo mantuvo un nivel adecuado de estabilidad y no mostró patrones de sesgo. Los residuales no presentaron estructuras que indican omisión de variables relevantes ni sobreajuste, lo que respalda la validez de la aproximación del modelo.

Capítulo IV

4. Análisis y Discusión de Resultados

Este apartado presenta los resultados del modelo de regresión supervisada diseñado para estimar el monto exportado esperado condicionado a que la empresa exporte, es decir, $E(monto \mid exporta, X)$. El análisis se realiza exclusivamente sobre el subconjunto de empresas exportadoras, en coherencia con la formulación teórica del modelo dual, evitando distorsiones asociadas a la presencia de ceros estructurales.

4.1 Resumen Ejecutivo de los Resultados

Antes de profundizar en el análisis detallado de cada componente del modelo, se presenta una síntesis de las métricas más relevantes obtenidas en el conjunto de prueba. Esta tabla resume el desempeño del clasificador, del regresor y del modelo combinado, tanto en su versión no ponderada como ponderada por el factor de expansión muestral.

Tabla 5.

Resumen de Métricas del Modelo Dual (Conjunto de Prueba)

Componente	Métrica	Valor
Clasificación (XGBoost)	ROC_AUC	0.930
Clasificación (XGBoost)	Exactitud (Accuracy)	0.891
Clasificación (XGBoost)	Precisión	0.829
Clasificación (XGBoost)	Recall	0.675
Clasificación (XGBoost)	F1-Score	0.744
Clasificación ponderada (f_exp)	ROC_AUC ponderado	0.927
Regresión (Random Forest)	R^2 (escala original)	0.375
Regresión (Random Forest)	R^2 (escala logarítmica)	0.700
Regresión (Random Forest)	MAE	USD 5.3 millones
Regresión (Random Forest)	RMSE	USD 43.8 millones
Combinación P × E	MAE	USD 1.36 millones
Combinación P × E ponderada (f_exp)	MAE ponderado	USD 420,775

Estos resultados evidencian que el modelo dual logra un desempeño sólido tanto en la identificación de empresas exportadoras como en la estimación del monto exportado, y que la combinación de ambos componentes reduce significativamente el error promedio al considerar toda la población empresarial.

4.2 Resultados del Clasificador

El modelo de clasificación basado en XGBoost presenta un desempeño elevado en la identificación de empresas exportadoras. El valor de ROC AUC de 0.93 indica una alta capacidad discriminativa, es decir, el modelo es capaz de diferenciar correctamente entre empresas exportadoras y no exportadoras en una amplia gama de umbrales de decisión.

La exactitud global del modelo alcanza aproximadamente el 89%, mientras que la precisión y el recall muestran un balance razonable considerando la naturaleza desbalanceada del problema. En particular, el recall del 67.5% refleja la capacidad del modelo para identificar una proporción significativa de exportadores reales, lo cual resulta especialmente relevante en aplicaciones de política pública donde el costo de omitir empresas exportadoras puede ser elevado.

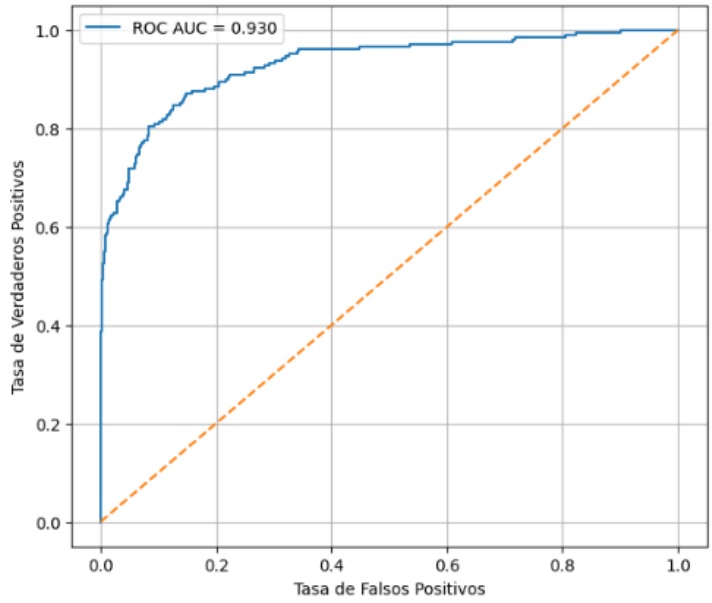
Al incorporar el factor de expansión muestral en la evaluación, el ROC AUC ponderado se incrementa hasta aproximadamente 0.93, lo que indica que el modelo no solo discrimina bien a nivel muestral, sino que también mantiene su capacidad predictiva cuando se pondera por la representatividad poblacional de cada empresa.

Este resultado es particularmente importante en el contexto de encuestas estructurales, ya que confirma que el modelo no está sesgado hacia empresas sobrerrepresentadas en la muestra y que su desempeño es consistente con la estructura real del tejido empresarial ecuatoriano.

4.2.1 Curva ROC

La curva ROC permite evaluar la capacidad discriminativa del modelo para distinguir entre empresas exportadoras y no exportadoras a lo largo de todos los posibles umbrales de decisión. A diferencia de métricas puntuales como la exactitud, esta curva resume el comportamiento global del clasificador frente al trade-off entre verdaderos positivos y falsos positivos, siendo especialmente útil en contextos de clases desbalanceadas, como el presente estudio.

Figura 7.
Curva ROC - Clasificador XGBoost



La curva ROC se mantiene claramente por encima de la diagonal de referencia, lo que confirma que el modelo posee una elevada capacidad discriminativa. El valor de AUC cercano a 0.93 indica que, al seleccionar aleatoriamente una empresa exportadora y una no exportadora, el modelo asigna una probabilidad mayor a la empresa exportadora en más del 93% de los casos.

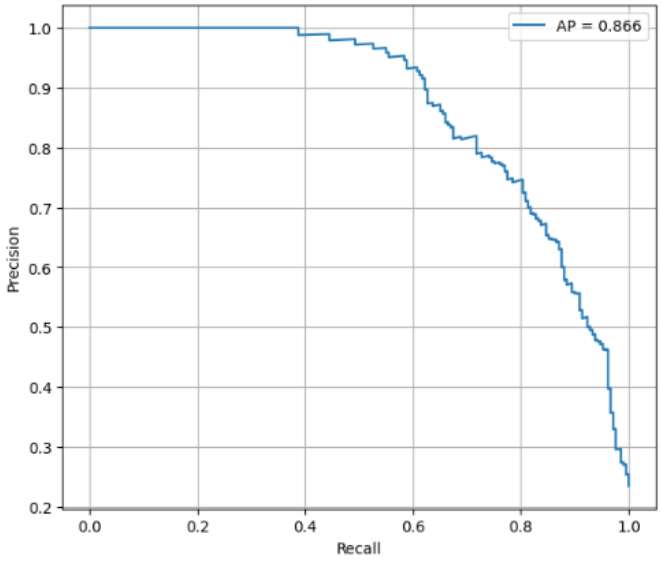
Desde una perspectiva económica, este resultado sugiere que las variables estructurales incluidas en el modelo capturan adecuadamente los determinantes de la

inserción exportadora, tales como tamaño empresarial, nivel de producción y características sectoriales.

4.2.2 Curva Precision–Recall

Dado el desbalance de clases existente en la variable objetivo, la curva Precision–Recall resulta especialmente informativa, ya que enfatiza el desempeño del modelo sobre la clase positiva (empresas exportadoras), permitiendo evaluar el equilibrio entre la precisión y la cobertura del modelo.

Figura 8.
Curva Precisión-Recall - Clasificación

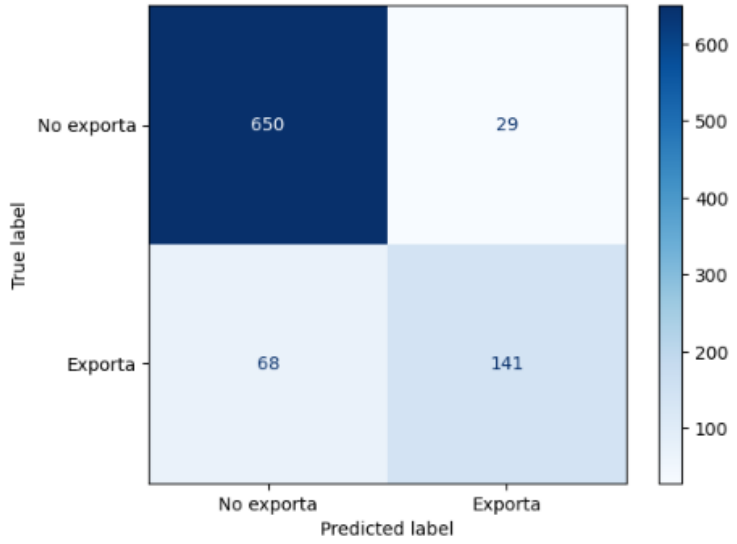


La curva Precision–Recall muestra que el modelo mantiene niveles elevados de precisión incluso cuando el recall se incrementa, lo cual evidencia que el aumento en la detección de empresas exportadoras no implica un crecimiento desproporcionado de falsos positivos. Este comportamiento es deseable en aplicaciones donde los recursos para intervenciones o análisis posteriores son limitados y se requiere priorizar casos con alta probabilidad real de exportación.

4.2.3 Matriz de Confusión

La matriz de confusión permite analizar de forma directa los errores de clasificación, identificando la frecuencia de falsos positivos y falsos negativos, lo cual resulta clave para interpretar el impacto operativo del modelo.

Figura 9.
Matriz de Confusión



La matriz de confusión evidencia que el modelo clasifica correctamente a la mayoría de las empresas no exportadoras, manteniendo una tasa controlada de falsos positivos. Aunque existe una proporción de falsos negativos, este resultado es coherente con la complejidad del fenómeno exportador, donde algunas empresas presentan características similares a las exportadoras sin necesariamente participar en mercados internacionales.

Desde el punto de vista de la política pública este patrón implica que el modelo es mesurado al identificar a empresas exportadoras, generando así una reducción del riesgo al asignar recursos a empresas sin capacidad real de exportación.

4.3 Resultados del Modelo de Regresión

Este apartado presenta los resultados del modelo de regresión supervisado diseñado para estimar el monto exportado esperado condicionado a que la empresa exporte, es decir,

$E(monto \mid exporta, X)$. El análisis se realiza exclusivamente sobre el subconjunto de empresas exportadoras, en coherencia con la formulación teórica del modelo dual.

4.3.1 Métricas de Desempeño en el Conjunto de Prueba (Empresas Exportadoras)

En este subapartado se evalúa la capacidad predictiva del modelo de regresión en el conjunto de prueba, considerando únicamente observaciones con monto exportado positivo ($v1005 > 0$). Las métricas se calculan en la escala original del monto exportado, tras aplicar la transformación inversa $exp\,(\cdot) - 1$, con el fin de asegurar interpretabilidad económica directa. Se reportan las siguientes métricas:

i. **Error Absoluto Medio (MAE)**: mide el error promedio en dólares, siendo especialmente relevante para evaluar precisión operativa.

ii. **Raíz del Error Cuadrático Medio (RMSE)**: penaliza con mayor intensidad errores grandes, sensible a desviaciones extremas.

iii. **Coeficiente de Determinación (R^2)**: indica la proporción de variabilidad del monto exportado explicada por el modelo.

Tabla 6.
Métricas de Desempeño del Modelo de Regresión

Modelo	MAE (USD)	RMSE (USD)	R^2
Random Forest	5.317.424	43.785.499	0.375
XGBoost	5.858.578	43.960.530	0.370

Los resultados evidencian que ambos modelos logran explicar aproximadamente el 37% de la variabilidad total del monto exportado. El Random Forest presenta un menor MAE, razón por la cual fue seleccionado como modelo final para la estimación condicional.

4.3.2 Métricas Ponderadas por el Factor de Expansión (f_exp)

Dado que los datos provienen de una encuesta bajo un diseño muestral , las métricas tradicionales pueden subestimar errores en empresas con mayor representatividad poblacional. Por ello, se ajustan las métricas incorporando el factor de expansión como peso muestral. Este enfoque permite evaluar el desempeño del modelo apegado a datos poblacionales.

En este sentido, el análisis de errores ponderados determina que el modelo mantiene un comportamiento estable cuando se evalúa desde una perspectiva poblacional, lo cual es de suma importancia para la aplicación de política pública basada en información estadística robusta.

Tabla 7.
Métricas del Regresor Considerando Ponderación Muestral

Métrica	Valor
MAE ponderado (USD)	≈ 5.3 millones
RMSE ponderado (USD)	≈ 43.8 millones

Al analizar la tabla, se observa que el desempeño del modelo no se ve distorsionado por valores extremos y que la estrategia de normalización del factor de expansión aplicada en el entrenamiento contribuye a la estabilidad numérica del estimador.

4.3.3 Análisis de Residuos: Heterocedasticidad, Valores Atípicos y Winsorización

El análisis de residuos del modelo de regresión se realizó comparando los valores observados del monto exportado con las predicciones obtenidas en el conjunto de prueba, restringido únicamente a empresas exportadoras. Los estadísticos descriptivos de los residuos muestran una media cercana a cero, lo que indica ausencia de sesgo en las predicciones, mientras que la dispersión aumenta a medida que incrementa el rango del monto exportado.

Este patrón se presenta con la presencia de heterocedasticidad, mayoritariamente influenciados donde las empresas de mayor tamaño presentan una mayor variabilidad en sus resultados. De la misma manera, se observa que los errores más determinantes se concentran en la cola superior de la distribución (cola derecha), lo cual, es esperable dado el reducido número de empresas con exportaciones representativamente altas. La ausencia de patrones sistemáticos en los residuos sugiere que el modelo obtiene adecuadamente la relación central entre las variables explicativas y el monto exportado, sin embargo, persiste incertidumbre en la predicción de valores extremos, lo cual, es un aspecto que refuerza la modelización del enfoque dual adoptado en el estudio de investigación.

Por último, se evaluó el impacto de la winsorización aplicada exclusivamente en el conjunto de entrenamiento con el fin de analizar su efecto sobre la estabilidad y capacidad predictiva del modelo.

La winsorización se justifica por la existencia de valores extremadamente altos de exportación que, de no mantener un tratamiento adecuado de los datos pueden deteriorar la capacidad de generalización.

Tabla 8.

Desempeño del Regresor Según Nivel de Winsorización

Nivel de winsorización	MAE (USD)	RMSE (USD)	R^2 (escala original)
p = 0.99	5.317.424,05	43.785.499,48	0.3754

Los resultados muestran que la winsorización representa una mejora en el equilibrio entre sesgo y varianza, disminuyendo la sensibilidad del modelo por valores extremos sin eliminar el sesgo presente en la cola superior. Esta evidencia respalda la elección de un percentil de winsorización alto (p = 0.99) como un compromiso adecuado entre robustez estadística y preservación de la información económica relevante.

4.3.4 Resultados de la Combinación Expected = P·E

El enfoque dual propuesto en esta investigación culmina en la estimación del valor esperado de exportaciones mediante la combinación de dos componentes: la probabilidad de que una empresa exporte y el monto esperado condicionado a que dicha exportación ocurra.

Esta estrategia permite abordar de manera analítica el fenómeno exportador, caracterizado por una elevada proporción de ceros estructurales (empresas no exportadoras) y una distribución altamente asimétrica en los valores positivos (grandes exportadores).

Desde una perspectiva metodológica, la variable $expected = P(exporta|X) \cdot E(monto|exporta, X)$ constituye una estimación del comportamiento exportador, válida tanto para empresas exportadoras como no exportadoras, lo cual habilita su evaluación sobre la totalidad de la muestra de test.

4.3.5 Métricas del Modelo Combinado

La evaluación del modelo combinado se realizó sobre el conjunto de test completo, considerando como variable objetivo los valores observados de exportaciones (v1005), incluyendo explícitamente los ceros correspondientes a empresas no exportadoras. Este enfoque permite medir la capacidad real del modelo para determinar el comportamiento del fenómeno exportador.

Los resultados obtenidos muestran que el error absoluto medio (MAE) del valor esperado se sitúa en USD 1,356,119, mientras que el error cuadrático medio (RMSE) alcanza USD 21,457,964. La diferencia entre ambas métricas responde a la presencia de empresas con volúmenes de exportación excepcionalmente altos.

Tabla 9.
Métricas del Modelo Combinado Expected en el Conjunto de Test

Métrica	Valor
MAE (USD)	1,356,119.68
RMSE (USD)	21,457,964.49

Los resultados mostrados en la tabla evidencian que el modelo logra una aproximación adecuada del valor promedio de las exportaciones empresariales, manteniendo errores razonables en relación con la escala económica del fenómeno exportador.

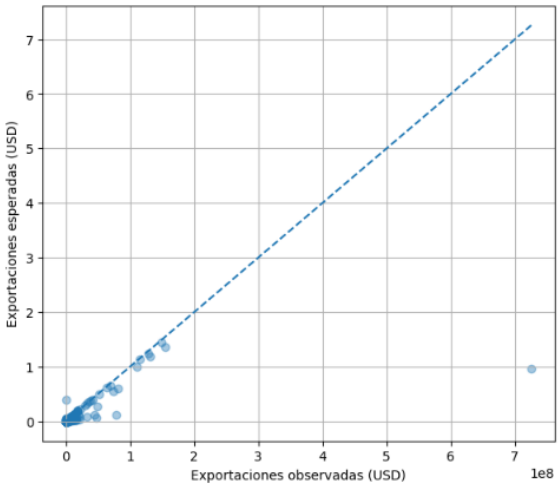
Dado el diseño muestral complejo de la Encuesta Estructural Empresarial (ENESEM), se incorporó el factor de expansión (f_exp) para evaluar el desempeño del modelo desde una perspectiva poblacional, como se ha mencionado en la investigación. Al ponderar el MAE mediante este factor, el error promedio se reduce a USD 420,775, lo cual indica que el modelo presenta un ajuste superior en las observaciones de mayor peso poblacional.

Este resultado es determinante para encuestas de estadística oficial, ya que sugiere que el modelo es consistente no solo a nivel muestral, sino también al proyectarse sobre el universo empresarial. En consecuencia, la suma ponderada del valor esperado ($total\_est = \Sigma\ expected_i \cdot f\_exp_i$) constituye una estimación coherente del total poblacional de exportaciones empresariales.

El análisis gráfico de la relación entre exportaciones observadas y valores esperados permite evaluar visualmente la calidad del ajuste del modelo, así como identificar patrones sistemáticos de sobreestimación o subestimación.

Ilustración 10.

Dispersión Entre Exportaciones Observadas y Esperadas



El gráfico evidencia una alta concentración de observaciones en valores bajos, reflejando la estructura real de la población empresarial. La cercanía de los puntos a la diagonal confirma un buen ajuste promedio del modelo. En los valores más altos se observa mayor dispersión, lo cual es consistente con la elevada heterogeneidad de las grandes empresas exportadoras.

4.4 Interpretabilidad del Modelo Dual

El enfoque dual permite distinguir explícitamente entre los determinantes de la probabilidad de exportar y aquellos que explican la magnitud del monto exportado. Esta separación aporta una ventaja interpretativa sustancial frente a modelos tradicionales de regresión simple.

4.4.1 Importancia de Variables en el Clasificador y el Regresor

El análisis de importancia global de variables evidencia una diferenciación clara entre los factores que determinan la probabilidad de que una empresa exporte y aquellos que explican el monto exportado condicional a la participación en mercados externos (empresa exportadora).

En el modelo clasificador, las variables con mayor peso están asociadas principalmente a características estructurales de la empresa, tales como su tamaño, sector económico y trayectoria productiva, las cuales influyen directamente en la capacidad de acceso al comercio internacional.

En contraposición, el modelo regresor asigna mayor relevancia a variables vinculadas con la escala productiva, financiera y operativa, lo que sugiere que, una vez que la empresa ha superado la barrera de entrada al mercado exportador, el volumen exportado depende fundamentalmente de su capacidad instalada, nivel de ventas, inversión en capital fijo y estructura de costos laborales.

La comparación entre ambos modelos confirma que los determinantes de la decisión de exportar no coinciden plenamente con los determinantes del volumen exportado, lo cual justificó empíricamente el uso del enfoque dual implementado en esta investigación.

Tabla 10.

Determinantes de los Modelos de Clasificación y Regresión

Clasificador – Probabilidad de exportar (P)	**Regresor – Monto exportado condicional (E)**
Tamaño de la empresa (cod_tamano)	Ventas brutas de producción (Vbp)
Sector económico (cod_ciiu2d)	Ventas netas (Vns)
Antigüedad del RUC (anio_ruc_dis)	Formación bruta de capital fijo (fbk)
Provincia	Remuneraciones totales (totremun)
Empleo total (totalpeoc)	Activos productivos (v4001)

La tabla muestra que el clasificador se apoya en variables que capturan estructura, localización y tamaño, las cuales son determinantes para la inserción en mercados externos. Por su parte, el regresor enfatiza variables directamente relacionadas con la capacidad productiva y financiera, lo que explica diferencias en el volumen exportado entre empresas que ya participan en el comercio internacional.

Capítulo V

5.1 Conclusiones

En esta investigación se logró el objetivo de desarrollar un modelo de predicción dual partiendo del enfoque econométrico two-part model el cual es principalmente utilizado en eventos o situaciones donde la probabilidad de participación está condicionada por factores diferentes a la magnitud de la participación, este enfoque presenta una arquitectura de doble etapa siendo la primera una clasificación de empresas exportadoras y no exportadoras y una segunda etapa en la que se predice el valor de la exportación. En conjunto ambas técnicas dieron como resultado una predicción más acertada del monto de exportación, este es un indicador muy significativo en el análisis económico pues muestra el potencial exportador de las empresas y cuáles son las características que más influyen en este proceso.

Para el modelo se utilizó una base de datos estructurada por el Instituto Nacional de Estadísticas y Censos (INEC) con 4.436 registros, debido al complejo diseño muestral que se maneja en la encuesta los datos fueron ponderados con el factor de expansión durante los procesos de entrenamiento y evaluación del modelo, esta ponderación fue indispensable para obtener predicciones representativas y generalizables a nivel poblacional. Además se realizó una estratificación en los conjuntos de entrenamiento y prueba para mantener consistencia en los datos y evitar fuga de información.

En la primera etapa de clasificación el algoritmo XGBoost mostró el mejor rendimiento. La métrica de evaluación ROC_AUC obtuvo un valor inicial de 0,93 lo que indica gran capacidad de clasificación entre empresas exportadoras y no exportadoras a pesar del desbalance de clases, con la ponderación del factor de expansión esta métrica incluso fue más alta lo que demuestra la importancia de ajustar los datos a la estructura poblacional.

En la segunda etapa de regresión el algoritmo RandomForestRegressor tuvo el mejor desempeño, la estabilidad del modelo incementó con la transformación logarítimca alcanzando un R^2_log de aproximadamente 0,70, es decir la mayor parte de la variabilidad de los datos se pudo explicar con el modelo y sus variables predictoras, sin embargo en la escala original el rendimiento del modelo se vio afectada por valores extremos (mega-exportadores), este es un comportamiento habitual en las economías a gran escala, sin embargo esto se mitigó con la transformación logarítmica.

En conjunto la probabilidad de que una empresa exporte por el valor predicho de exportación dio inicialmente un valor de MAE de 1.36 millones, sin embargo con la ponderación del factor de expansión este valor disminuyó a 421mil, reafirmando la importancia de este factor para la extrapolación de los resultados y la disminución de errores.

Con estos resultados se concluye que el enfoque adoptado para este proyecto es robusto metodológicamente, los resultados son coherentes con la literatura, tiene relevancia estadística y por lo tanto también es útil operativamente. El modelo identifica y captura por un lado patrones relevantes que determinan si una empresa tiene potencial exportador y también el monto de exportación. En este sentido podría ser una herramienta de apoyo en la toma de decisiones, así como en el planteamiento de políticas públicas que referencien al sector exportador.

Respecto a las limitaciones del modelo que si bien no comprometen su validez predictiva deben ser considerados en la interpretación y extrapolación de los resultados. Se encontraron los siguientes factores: la principal limitante es la fuerte heterogeneidad de las empresas de los diferentes sectores económicos, también la colinealidad de variables económicas y productivas, los valores extremos que distorsionan la distribución y la

naturaleza de los modelos basados en árboles que no necesariamente debe ser interpretada como causalidad directa.

5.2 Recomendaciones

La winsorización al 0.99 tuvo una influencia significativa en la regresión, se podría realizar un análisis de sensibilidad con otros percentiles para determinar un equilibrio entre la robustez del indicador y la conservación de la estructura de datos para no perder variaciones relevantes de la información. Además para reducir el impacto de los valores extremos se podría utilizar alternativas como Huber o Regresión por cuantiles ya que sus funciones se basan en pérdidas robustas.

Debido a la heterogeneidad de los diferentes sectores exportadores se podría considerar evaluar modelos por sectores para captar mejores patrones que puedan brindar más información e interpretabilidad. En caso de aumentar el número de variables predictoras que acentúen la colinealidad interna de las variables se podría implementar técnicas como la reducción de dimensionalidad o un análisis de componentes principales para incrementar la estabilidad del modelo.

Para aplicaciones del modelo en áreas más sensibles que requieran mayor nivel de fiabilidad se debería implementar análisis de cuantificación de incertidumbre, con intervalos de predicción usando metodologías bayesianas o técnicas como el Bootstrap mediante las cuales se pueda evaluar el nivel de variabilidad de las predicciones.

Respecto a la ponderación con el factor de expansión se recomienda mantenerlo durante el entrenamiento y evaluación del modelo ya que se evidenció una mejora relevante en las métricas del modelo y además los resultados puedes ser extrapolados respetando la

estructura poblacional y evitando sesgos inherentes a la muestra ya que la base de datos proviene de una encuesta.

Construir un sistema de gobernanza para que asegure la reproducibilidad del modelo, esto implica guardar el procesador de datos, los límites de winsorización, los hiperparámetros óptimos, los modelos entrenados y en general los artefactos necesarios para volver a construir el modelo predictivo.

La actualización periódica del modelo también es muy importante, ya que al abordar un tema económico tan complejo como es el entendimiento del sector exportador es necesario mantener constantes evaluaciones y actualizaciones de los componentes del modelo, como los límites de la winsorización, el ajuste de las probabilidades en el modelo de clasificación o con la incorporación de nuevas variables que más adelante podrían ser relevantes y que evidencien cambios en el ámbito económico.

A manera de conclusión, el modelo dual desarrollado en este proyecto plantea una herramienta de predicción sólida y robusta a nivel metodológico que permite analizar y profundizar en el comportamiento de las empresas exportadoras, sus aplicaciones son amplias especialmente en la práctica económica. El modelo de clasificación XGBoost y el modelo de regresión RandomForest en conjunto muestran una aproximación balanceada entre la capacidad de clasificación, la generalización de los resultados y la aplicabilidad de la información. A pesar de que aún existen oportunidades de mejora asociados principalmente con la composición de los datos con presencia de valores extremos y la heterogeneidad, el sistema muestra resultados analíticamente significativos y coherentes con la teoría.

Bibliografía

Aldrin, C., Aviles, E., Baque, E., & Muñiz, F. (2024). El papel de la analítica predictiva en la anticipación de cambios en el entorno empresarial. *Ciencia y Desarrollo*, 27(2), 43–56. http://revistas.uap.edu.pe/ojs/index.php/CYD/index

Álvarez, R., & López, R. (2005). Exporting and performance: Evidence from Chilean firms. *Empirical Economics*, 30(3), 589–607. https://doi.org/10.1007/s00181-005-0256-7

Banco Central del Ecuador. (2020). *Informe de comercio exterior del Ecuador*. https://www.bce.fin.ec

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. https://fairmlbook.org

Bernard, A. B., & Jensen, J. B. (1999). Exceptional exporter performance: Cause, effect, or both? *Journal of International Economics*, 47(1), 1–25. https://doi.org/10.1016/S0022-1996(98)00027-0

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5), 829–844. https://doi.org/10.2307/1909582

Cruz, E. S., & Brito, R. (2023). Modelo de correlación entre exportaciones e importaciones del Ecuador entre los años 2010 y 2019. *Ciencia Latina Revista Científica*. https://doi.org/10.37811/cl_rcm.v7i4.7339

Estévez, M. (2017). *Modelos econométricos*. Inteligencia Analítica. https://inteligencia-analitica.com/modelos-econometricos/

Garzón, J., & Díaz, G. (2023). Mínimos cuadrados parciales y su aplicación en la identificación de competitividad en empresas exportadoras de banano de la ciudad de Machala. *Revista Científica Portal de la Ciencia*, 221–231.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161. https://doi.org/10.2307/1912352

Helpman, E. (2008). *The mystery of economic growth*. Harvard University Press.

Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons.

Investigación y Marketing. (2023). El valor de las muestras representativas. *Investigación y Marketing*, (155). https://ia-espana.org/wp-content/uploads/2023/05/Revista-155-IM.pdf

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

López, R. (2009). Exportaciones y desempeño empresarial: Evidencia empírica y desafíos metodológicos. *Revista de Economía Aplicada*, 17(49), 5–34.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. En *Advances in Neural Information Processing Systems* (Vol. 30). https://proceedings.neurips.cc

Martínez, F., & Cobos, C. (2025). Revisión sistémica de las herramientas de inteligencia artificial explicable usadas en métodos de ensamble. *Revista Facultad de Ingeniería*, 33(70), 1–20. https://doi.org/10.19053/01211129.v33.n70.2024.18078

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6), 1695–1725. https://doi.org/10.1111/1468-0262.00467

OECD. (2021). *Artificial intelligence, machine learning and policy*. OECD Publishing. https://www.oecd.org

Ordoñez, L., Pinos, L., & García, P. (2020). Elasticidad-renta del comercio bilateral mediante el modelo gravitacional: Caso Ecuador. *Revista Economía y Política*, 15(30), 1–30. https://www.redalyc.org/journal/5711/571162031007/html/

Orellana, J. (2025). Análisis de riesgo bursátil y modelo de predicción de las acciones de la empresa Holcim S.A. *Revista Decisión Gerencial*, 4(9), 54–73. https://doi.org/10.26871/rdg.v4i9.64

Palacios, Á., & Josué, C. (2023). *The impact of trade on intra-industry reallocations and aggregate industry productivity*. Universidad Católica Santiago de Guayaquil. http://repositorio.ucsg.edu.ec/handle/3317/21443

Paredes, E., Vinces, L., Castro, D., & Llerena, R. (2025). El impacto de las exportaciones en el crecimiento económico de Ecuador (2000–2023): Un análisis empírico basado en el modelo de Feder. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, 6(2), 35–46. https://doi.org/10.56712/latam.v6i2.3603

Pinos, L., Cevallos, E., Abril, X., & Pauta, J. (2025). Determinantes de exportación del sector fabricante ecuatoriano 2018–2021: Un análisis a nivel empresa. *Economía y Negocios*, 16(1), 57–75. https://www.redalyc.org/journal/6955/695580042004/html/

Rudziński, F., Guillén, M., & Nguyen, Q. (2022). Machine learning for international trade and firm export survival. *Journal of Economic Surveys*, 36(2), 456–489. https://doi.org/10.1111/joes.12455

Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

Toaza, P., & Rivera, A. (2025). Análisis de la competitividad de las exportaciones ecuatorianas en el mercado global 2021–2023. *REVISTVR*, 1–20. https://revista.istvr.edu.ec/wp-content/uploads/2025/03/ANALISIS-DE-LA-COMPETITIVIDAD-DE-LAS-EXPORTACIONES-ECUATORIANAS-EN-EL-MERCADO-GLOBAL-2021-2023.pdf

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91. https://doi.org/10.1186/1471-2105-7-91

Wagner, J. (2007). Exports and productivity: A survey of the evidence from firm-level data. *The World Economy*, 30(1), 60–82. https://doi.org/10.1111/j.1467-9701.2007.00872.x

Wooldridge, J. M. (2010). *Econometric analysis of cross-section and panel data* (2nd ed.). MIT Press.

Apéndice A

Diccionario de Variables

Tipo de Variable	Variable	Descripción
Cualitativa	provincia	Provincia donde opera la empresa (localización geográfica)
	cod_tamano	Tamaño de la empresa según categorías INEC
	des_tamano	Descripción del tamaño de empresa
	cod_ciiu2d	Código CIIU a 2 dígitos (rama de actividad)
	des_ciiu2d	Descripción del CIIU
	cod_sector	Clasificación sectorial
	des_sector	Descripción del sector económico
Cuantitativa	anio_ruc_dis	Año de inscripción o actualización del RUC
	capital_social_nac_privado	Capital social nacional de origen privado
	capital_social_nac_publico	Capital social nacional de origen público
	capital_social_ext_privado	Capital social extranjero privado
	capital_social_ext_publico	Capital social extranjero público
	v1001	Ventas totales anuales
	v1002	Ventas locales
	v1003	Ventas nacionales
	v1004	Ventas al exterior distintas a exportaciones monetizadas
	v2001	Costos de producción
	v2002	Gastos operativos
	Vbp	Valor bruto de la producción
	Vbc	Valor bruto de la comercialización
	Vns	Valor neto de la producción
	v2006	Costos administrativos
	v4001	Gastos financieros
	v1208	Inversión en innovación
	totalpeoc	Total de personal ocupado
	totremun	Total de remuneraciones pagadas
	totadquisi	Valor total de adquisiciones
	fbk	Formación bruta de capital
	fbkf_1	Detalle de formación bruta de capital físico
	cant_ener	Cantidad de energía consumida
	cant_agua	Consumo anual de agua
	v1125	Compras de insumos nacionales
	v1126	Compras de insumos importados
	adqvnv	Total de adquisiciones varias

Apéndice B

Código Fuente del Modelo Dual

```
# ---------- LIBRERÍAS PRINCIPALES ----------
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
from sklearn.metrics import roc_auc_score, mean_absolute_error
from joblib import dump
# ---------- PARÁMETROS GENERALES ----------
RANDOM_STATE = 42      # Reproducibilidad
WINSOR_P = 0.99        # Winsorización (solo TRAIN)
# ---------- FUNCIÓN AUXILIAR ----------
def clean_numeric_series(s):
    """
    Limpieza robusta de variables numéricas almacenadas como texto.
    """
    s = s.astype(str).str.replace(",", ".", regex=False)
    return pd.to_numeric(s, errors="coerce")
```

```
# ---------- CARGA DE DATOS ----------
df = pd.read_csv("INEC_Encuesta_Estructural_Empresarial_2023.csv", sep=";")
# Variable binaria: empresa exportadora
df["v1005_numeric"] = clean_numeric_series(df["v1005"])
df["exportador"] = (df["v1005_numeric"].fillna(0) > 0).astype(int)
# ---------- SELECCIÓN DE VARIABLES ----------
features = [
    "provincia","cod_tamano","cod_sector","v1001","v1002",
    "Vbp","Vns","totalpeoc","totremun","fbk"
]
X = df[features]
y = df["exportador"]
# ---------- TRAIN / TEST (SIN LEAKAGE) ----------
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=RANDOM_STATE
)
# ---------- FACTOR DE EXPANSIÓN ----------
if "f_exp" in df.columns:
    w_train          =          clean_numeric_series(df.loc[X_train.index,
"f_exp"]).fillna(1)
else:
    w_train = None

# ---------- PREPROCESAMIENTO ----------
```

```
num_features = ["v1001","v1002","Vbp","Vns","totalpeoc","totremun","fbk"]
cat_features = ["provincia","cod_tamano","cod_sector"]
preprocessor = ColumnTransformer([
    ("num", Pipeline([
        ("imp", SimpleImputer(strategy="median")),
        ("sc", StandardScaler())
    ]), num_features),
    ("cat", Pipeline([
        ("imp", SimpleImputer(strategy="constant", fill_value="missing")),
        ("oh", OneHotEncoder(handle_unknown="ignore"))
    ]), cat_features)
])

# ---------- ETAPA 1: CLASIFICACIÓN ----------
clf = Pipeline([
    ("pre", preprocessor),
    ("rf", RandomForestClassifier(
        n_estimators=200, random_state=RANDOM_STATE))
])
clf.fit(X_train, y_train, rf__sample_weight=w_train)
p_export = clf.predict_proba(X_test)[:, 1]
roc = roc_auc_score(y_test, p_export)
# ---------- ETAPA 2: REGRESIÓN (SOLO EXPORTADORES TRAIN) ----------
train_exp = df.loc[X_train.index]
train_exp = train_exp[train_exp["v1005_numeric"] > 0]
```

```
cap = train_exp["v1005_numeric"].quantile(WINSOR_P)
y_r = np.log1p(train_exp["v1005_numeric"].clip(upper=cap))
X_r = X_train.loc[train_exp.index]
reg = Pipeline([
    ("pre", preprocessor),
    ("rf", RandomForestRegressor(
        n_estimators=300, random_state=RANDOM_STATE))])
reg.fit(X_r, y_r)
# ---------- COMBINACIÓN EXPECTED ----------
pred_amount = np.expm1(reg.predict(X_test))
expected = p_export * pred_amount
mae_expected = mean_absolute_error(
    df.loc[X_test.index, "v1005_numeric"].fillna(0),
    expected
)
# ---------- GUARDADO ----------
dump(clf, "classifier_best.joblib")
dump(reg, "regressor_best.joblib")

print("ROC-AUC:", roc)
print("MAE expected:", mae_expected)
```