**Pontificia Universidad Católica del Ecuador**

**Facultad de Hábitat, Infraestructura y Creatividad**

**Maestría en Sistemas de Información mención Data Science**


Trabajo previo para la obtención del título de Magister en Sistemas de Información mención Data Science

## Tema:

# Predicción probabilística del tipo de aforo aduanero en importaciones courier a Ecuador: Clasificación multiclase (Automático, Documental, Físico Intrusivo y Físico No Intrusivo-RX) con aprendizaje automático en Python

## Autor:

Kevin Andrés Romero Arteaga

## Tutor:

Damián Aníbal Nicolalde Rodríguez Mg.

Quito - febrero 2026

**Tabla de Contenidos:**

**Índice de Ilustraciones:**



**Índice de Tablas:**

## 1. Título del proyecto

Predicción probabilística del tipo de aforo aduanero (automático, documental, físico intrusivo y físico no intrusivo-RX) en importaciones courier a Ecuador mediante modelos de aprendizaje automático en Python.

## 2. Resumen

Este trabajo propone el diseño y evaluación de un modelo predictivo que estime, para cada paquete de importación courier en Ecuador, la probabilidad de asignación a cuatro clases de aforo: automático, documental, físico intrusivo y físico no intrusivo (RX). El problema se formula como una clasificación multiclase con severo desbalance, por lo que se emplean métricas robustas (macro-F1, balanced accuracy) y medidas específicas para eventos raros como curvas Precision-Recall (Saito & Rehmsmeier, 2015). Dado que el objetivo operativo requiere probabilidades confiables, se incorpora evaluación de calibración probabilística y el uso de puntuaciones probabilísticas (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017).

La metodología se alinea con CRISP-DM, emplea partición temporal para evitar fuga de información y combina modelos supervisados con segmentación no supervisada para descubrir perfiles latentes de riesgo (He & Garcia, 2009; OSCE, s. f.). Se utilizan datos históricos anonimizados del courier (N=128,484), donde la clase automática representa ~95.9% de los casos. Los resultados preliminares muestran que la etapa de detección de casos no automáticos alcanza ROC-AUC=0.731 y PR-AUC=0.186, con umbral operativo para recall≈0.80 y precisión≈0.12, priorizando la capacidad de alerta temprana. Se discuten implicaciones operativas (gestión de tiempos y comunicación al cliente), aportes académicos (identificación de factores asociados a mayor control) y límites del estudio (no causalidad y dependencia del conjunto de datos del courier).

## 3. Introducción

El despacho aduanero es una fase crítica de la cadena logística del comercio exterior, pues condiciona los tiempos de liberación, costos operativos y la experiencia del usuario final. En Ecuador, el Servicio Nacional de Aduana del Ecuador (SENAE) aplica modalidades de aforo que varían en intensidad de control. De acuerdo con la guía oficial de trámite de SENAE, para mercancías que requieren Declaración Aduanera se emplean modalidades como aforo automático, documental y físico, y se indica explícitamente que estas se seleccionan en función de perfiles de riesgo (SENAE, 2025a).

En paralelo, SENAE ha publicado manuales específicos para la ejecución de cada modalidad. Por ejemplo, el manual del canal de aforo documental describe el flujo de revisión mediante el sistema aduanero ECUAPASS (SENAE, 2024a), mientras que el manual del canal de aforo físico intrusivo detalla la inspección física y la necesidad de revisión documental asociada (SENAE, 2025b). Asimismo, el aforo físico no intrusivo (RX) dispone de un manual operativo propio, enfocado en inspección mediante tecnología no intrusiva (SENAE, 2024b). Sin embargo, estos documentos se concentran en el procedimiento, y no en la divulgación de reglas cuantitativas o ponderaciones exactas de perfilamiento, lo cual es coherente con

buenas prácticas internacionales de gestión de riesgo en aduanas, establecidas por la Organización Mundial de Aduanas, o WCO por sus siglas en inglés (WCO, s. f.-a).

En el régimen courier, la capacidad de anticipar la probabilidad de un canal de aforo más intensivo es valiosa para planificación operativa y comunicación al cliente: permite preparar documentación, estimar tiempos y priorizar recursos. Desde el punto de vista académico, el problema es relevante porque integra analítica predictiva en un contexto regulatorio donde el criterio exacto de selectividad no es público, pero existen resultados históricos observables. Por ello, este estudio busca (i) estimar probabilidades por clase de aforo y (ii) identificar factores observables asociados al incremento de control, reconociendo que el enfoque es predictivo-asociativo y no causal.

## 4. Objetivos de la investigación

### 4.1. Objetivo general:

Desarrollar y evaluar un modelo de aprendizaje automático que estime la probabilidad de asignación del tipo de aforo (automático, documental, físico intrusivo y físico no intrusivo-RX) para paquetes de importación courier en Ecuador, a partir de variables observables del paquete y su contexto operativo, generando un score probabilístico útil para planificación y comunicación al cliente.

### 4.2. Objetivos específicos:

- Consolidar y depurar la base histórica del courier con variables del paquete y el tipo de aforo asignado, generando diccionario de datos y reglas de calidad.
- Diseñar variables (feature engineering) enfocadas en características propias del paquete: valor, peso real, peso volumétrico, origen, categoría de envío, tipo de empaque, cantidad y contenido declarado.
- Entrenar y comparar modelos supervisados (multiclase y jerárquicos) considerando desbalance severo mediante ponderación de clases y métricas robustas.
- Evaluar calidad de probabilidades (calibración) y definir umbrales operativos para alertas tempranas (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017).
- Aplicar clustering para identificar perfiles latentes de paquetes y analizar su relación con la asignación de aforo (OSCE, s. f.).
- Documentar hallazgos y límites del estudio, incluyendo no causalidad, dependencia de datos del courier y consideraciones éticas.

## 5. Justificación del proyecto

El valor aplicado del proyecto radica en su utilidad operativa: un score probabilístico de aforo permite gestionar expectativas del cliente y optimizar recursos (p. ej., preparación documental, priorización de paquetes con alta probabilidad de control). En términos académicos, contribuye a la literatura de modelamiento de riesgo en contextos regulatorios al proponer una formulación multiclase desbalanceada y un enfoque de calibración probabilística para la toma informada de decisiones. A nivel de políticas de gestión de

riesgo, el estudio se alinea con la perspectiva de selectividad y focalización de controles recomendada por organismos internacionales (WCO, s. f.-a; OSCE, s. f.).

## 6. Marco teórico

### 6.1. Panorama introductorio: comercio exterior, aduanas y despacho

En comercio exterior, una importación es el ingreso de mercancías desde el exterior hacia el territorio aduanero de un país. Para que dichas mercancías puedan ser retiradas y circular legalmente, usualmente deben cumplirse formalidades aduaneras: presentación de una declaración, pago de tributos (cuando corresponda) y cumplimiento de controles y restricciones técnicas. Estas formalidades se realizan bajo el marco normativo y los sistemas informáticos de la administración aduanera.

En Ecuador, el Comité de Comercio Exterior (COMEX) define instrumentos de política comercial y arancelaria, mientras que el Servicio Nacional de Aduana del Ecuador (SENAE) ejecuta el control y la facilitación del despacho aduanero. Documentos recientes del COMEX sobre el régimen courier destacan principios constitucionales y del Código Orgánico de la Producción, Comercio e Inversiones (COPCI), y recuerdan obligaciones derivadas del Acuerdo sobre Facilitación del Comercio de la Organización Mundial de Comercio (OMC), especialmente en lo relativo a 'envíos urgentes' o expedidos (COMEX, 2025).

En términos operativos, las importaciones se apoyan en documentos y registros típicos. Entre los más comunes están: (i) documento de transporte (p. ej., guía aérea y/o guías courier), (ii) factura comercial o soporte del valor, (iii) declaración aduanera (normal o simplificada), y (iv) documentos de control previo cuando la mercancía lo requiere. En el régimen courier, SENAE describe explícitamente como requisito la transmisión de una Declaración Aduanera Simplificada (DAS) en el sistema ECUAPASS y la presentación de factura/soporte de valor y documento de transporte (SENAE, s. f.).

En un despacho aduanero, la administración busca equilibrar dos objetivos que están en tensión: facilitar el comercio legítimo (despacho rápido, electrónico, simplificado) y mantener el control (evitar fraude, subvaloración, contrabando, incumplimientos sanitarios o técnicos). La literatura y guías internacionales de gestión de riesgo en aduanas enfatizan precisamente este balance: no es eficiente inspeccionar físicamente todo, por lo que se aplican mecanismos de selectividad basados en riesgo para focalizar los controles en operaciones con mayor probabilidad de incumplimiento (WCO, s. f.-a).

Finalmente, el término 'aforo' se usa para referirse al tipo de control asignado a una declaración o envío: puede ser automático (sin intervención), documental (revisión de documentos) o físico (inspección de la mercancía). En este estudio se modelan cuatro clases: aforo automático, documental, físico intrusivo y físico no intrusivo (RX), pues en la operación courier la inspección mediante tecnología (RX) representa una modalidad distinta a la apertura/intrusión física del contenido (SENAE, 2024b).

### 6.1.1. Glosario de términos básicos referentes a aduanas y courier

- **Régimen aduanero:** conjunto de normas que definen cómo se declara y controla una operación (p. ej., importación, tránsito, courier).
- **Courier / mensajería acelerada:** modalidad para envíos expeditos de paquetes, con declaración simplificada y categorías con límites de peso/valor (SENAE, s. f.).
- **Manifiesto:** registro consolidado de envíos que arriban bajo una misma operación logística; en el dataset se representa por ManifestId y su fecha de creación ManifestCreationDate.
- **DAS (Declaración Aduanera Simplificada):** declaración utilizada por courier para transmitir información del envío en ECUAPASS (SENAE, s. f.).
- **DAI (Declaración Aduanera de Importación):** declaración estándar para importaciones fuera de los límites del régimen simplificado.
- **Aforo automático:** canal donde el sistema permite el avance sin revisión documental o física directa.
- **Aforo documental:** canal que exige revisión de documentos y consistencias declarativas (SENAE, 2024a).
- **Aforo físico intrusivo:** canal con inspección física (apertura/verificación) y contraste con documentos (SENAE, 2025b).
- **Aforo físico no intrusivo (RX):** canal con inspección mediante tecnología de rayos X u otra inspección no intrusiva (SENAE, 2024b).
- **Selectividad / perfil de riesgo:** reglas o modelos que determinan qué porcentaje de operaciones pasan a control intensivo, con base en indicadores de riesgo (WCO, s. f.-a; OSCE, s. f.).

## 6.2. Modalidades de aforo y operación documental en Ecuador

SENAE estructura el control del despacho mediante modalidades de aforo que varían en intensidad. La guía oficial del trámite de aforo físico indica que para mercancías con Declaración Aduanera se utilizan modalidades como aforo automático, documental o físico, seleccionadas con base en perfiles de riesgo (SENAE, 2025a). Esta afirmación provee el sustento institucional para formular el problema como un proceso de selectividad.

En el nivel procedimental, SENAE publica manuales que describen la ejecución de cada canal. El manual SENAE-MEE-2-2-011-V4 define el objetivo, alcance y secuencia operativa del canal documental mediante ECUAPASS (SENAE, 2024a). El manual SENAE-MEE-2-2-004-V7 describe el aforo físico intrusivo e indica que el técnico operador debe realizar, además de la revisión física, una revisión documental equivalente a la del canal documental (SENAE, 2025b). Finalmente, el manual SENAE-MEE-2-2-047-V2 detalla el canal físico no intrusivo (RX), orientado a inspección con tecnología sin apertura directa del paquete (SENAE, 2024b).

### 6.2.1. Régimen courier o mensajería acelerada

El régimen Courier, también denominado mensajería acelerada, está diseñado para facilitar el ingreso de paquetes mediante operadores logísticos especializados, con procedimientos simplificados y tiempos de despacho reducidos. En términos de política comercial, la lógica del régimen se asocia a la noción de 'envíos urgentes' o expedidos: permitir un levante rápido manteniendo el control aduanero, y admitiendo que la autoridad pueda exigir documentación adicional o limitar el beneficio según tipo de mercancía (COMEX, 2025).

En Ecuador, SENAE describe este régimen en su guía oficial de trámite para la autorización de ingreso de mercancías mediante mensajería acelerada. Allí se establece que el proceso se dirige a empresas courier (operadores de comercio exterior) registradas en el sistema informático aduanero ECUAPASS, y que para categorías distintas a la A se debe transmitir una Declaración Aduanera Simplificada (DAS) junto con factura/soporte de valor y documento de transporte (SENAE, s. f.).

Para comprender el problema de investigación, es importante notar que el régimen courier combina dos realidades: (i) una altísima frecuencia de envíos con baja unidad de valor/peso (paquetes), y (ii) la necesidad de control selectivo para impedir uso indebido con fines comerciales informales. Esta tensión aparece explícitamente en comunicaciones oficiales recientes: por ejemplo, el Ministerio de Producción, Comercio Exterior, Inversiones y Pesca (MPCEIP) informó que el número de paquetes de categoría B, también llamada 4x4, creció de forma acelerada y que se adoptaron medidas para desincentivar el uso indebido, manteniendo al mismo tiempo los límites del régimen (MPCEIP, 2025).

### 6.2.2. Categorías del régimen courier: límites, documentación y efectos en el despacho

El régimen Courier se organiza por categorías que establecen límites de peso, valor y tipos de mercancía, así como requisitos documentales y tratamiento tributario. Desde una perspectiva de ciencia de datos, estas categorías son relevantes porque condicionan el universo de mercancías y, por tanto, los patrones de riesgo observables en variables como valor declarado, peso, origen y contenido.

En la guía de SENAE se listan categorías con reglas y excepciones. Por ejemplo, se describen requisitos generales (DAS transmitida en ECUAPASS, factura/soporte de valor, documento de transporte, y documentos de control previo cuando aplique). También se indican particularidades como el registro de IMEI para celulares en la categoría C, y reglas especiales para textiles y calzado en la categoría D, incluyendo restricciones de control previo en envíos posteriores (SENAE, s. f.).

Un punto clave para operación y análisis es que, si un envío no cumple los límites establecidos para su categoría, la declaración puede ser rechazada y la mercancía se traslada a depósito temporal para cumplir formalidades aduaneras correspondientes. Este mecanismo convierte el cumplimiento de parámetros (peso/valor/tipo) en un potencial

indicador de riesgo y en una fuente de retrasos operativos, por lo que variables como valor declarado y pesos adquieren relevancia predictiva (SENAE, s. f.).

En 2025, el COMEX aprobó una reforma arancelaria que aplica una tarifa fija (USD 20 por paquete) a los 'Paquetes por Correos Rápidos (Mensajería Acelerada o Courier)' en una subpartida específica, reforzando la importancia económica del régimen y su crecimiento reciente (COMEX, 2025). En paralelo, el MPCEIP reportó cifras de crecimiento del régimen 4x4 y mencionó el mantenimiento del tope anual por usuario (USD 1.600), lo que muestra que el régimen no solo es logístico, sino también un instrumento de política comercial y fiscal (MPCEIP, 2025).

Las categorías descritas en la guía de SENAE son:

- **Categoría A:** envíos que siguen un procedimiento especial (incluye solicitud de inspección/autorización), sin transmisión de declaración aduanera como en otras categorías (SENAE, s. f.).
- **Categoría B (4x4):** orientada a envíos personales con límites de peso/valor por envío; ha sido objeto de ajustes y medidas de control por su crecimiento y riesgo de uso indebido (MPCEIP, 2025; COMEX, 2025).
- **Categoría C:** incluye disposiciones específicas (por ejemplo, celulares nuevos requieren registro de IMEI en la DAS) (SENAE, s. f.).
- **Categoría D:** contempla textiles y calzado bajo reglas particulares; la guía describe impuestos específicos por kilogramo/par y condiciones de control previo a partir de importaciones sucesivas (SENAE, s. f.).
- **Categoría E:** contempla medicinas u otras mercancías sensibles con requisitos sanitarios o médicos específicos (SENAE, s. f.).
- **Categoría G:** régimen especial vinculado a núcleo familiar de migrante, con límites y controles de seguimiento (SENAE, s. f.).

### 6.2.3. Flujo operativo del despacho courier y puntos donde ocurre el aforo

En términos prácticos, el despacho courier inicia con la llegada de los envíos y la creación/registro de manifiestos y guías asociadas. La guía de SENAE describe un flujo operativo que incluye: generar guía de distribución, ingresar a ECUAPASS y transmitir la Declaración Simplificada (Importación), y transmitir la declaración con código de régimen 91. Posteriormente, se revisan observaciones, se absuelven requerimientos y se determinan tributos cuando corresponda (SENAE, s. f.).

El tipo de aforo se asigna dentro del proceso de despacho como resultado de mecanismos de selectividad por riesgo. Desde la perspectiva de este trabajo, la etiqueta AforoType es un resultado observable de dicho proceso. La importancia del régimen courier es que, dado el volumen masivo de envíos, la administración necesita automatizar y reservar el control intensivo (documental/físico/RX) para una fracción pequeña de paquetes, por lo cual el aforo automático tiende a dominar en la distribución de clases observada.

El aforo físico no intrusivo (RX) representa una modalidad intermedia: incrementa el control sin requerir apertura de paquetes, apoyándose en tecnología de inspección. En la práctica, RX puede utilizarse para seleccionar casos para revisión posterior o para descartar rápidamente sospechas. En este sentido, modelar RX como una cuarta clase permite capturar diferencias de decisión entre inspección tecnológica y apertura intrusiva, y evita tratar el 'físico' como un bloque homogéneo (SENAE, 2024b).

#### 6.2.4. Implicaciones del régimen courier para analítica predictiva: datos, sesgos y señales

El régimen courier produce datos con características particulares. En primer lugar, la distribución del objetivo (tipo de aforo) suele ser extremadamente desbalanceada: la gran mayoría de envíos recibe aforo automático. En segundo lugar, variables del paquete (valor declarado, peso, origen, tipo y contenido) tienden a contener señal predictiva, pero esa señal puede ser tenue y se mezcla con decisiones de política comercial, restricciones por categoría y factores de carga operativa. Por ello, la evaluación debe centrarse en métricas robustas a desbalance y en probabilidades calibradas, más que en exactitud global.

Además, algunos campos de los sistemas operativos aparecen solo después de que se asigna el aforo (por ejemplo, el nombre del aforador). En ciencia de datos, incluir variables que son consecuencia del resultado en el entrenamiento crea 'fuga de información' (data leakage), inflando artificialmente el desempeño. Por ello, este trabajo explicita criterios de exclusión de variables posteriores al aforo y prioriza el uso de información disponible al momento del arribo o transmisión de la DAS, coherente con el objetivo operativo de anticipación.

#### 6.2.5. Hipótesis de factores propios del paquete

La literatura de gestión de riesgo aduanero señala que los indicadores de riesgo suelen combinar dimensiones del operador (cumplimiento histórico) con dimensiones de la mercancía y de la operación (origen, naturaleza, valor, consistencia documental, rutas). En escenarios donde no se dispone de toda la historia del operador, las variables del paquete actúan como extensiones útiles. Guías de selectividad y formación en gestión de riesgo enfatizan precisamente la combinación de indicadores para elevar o reducir el riesgo asignado (WCO, s. f.-a; OSCE, s. f.).

En un contexto courier, se espera que variables como valor declarado, peso real, peso volumétrico, país de origen y señales textuales del contenido influyan en la probabilidad de canal documental o físico. Por ejemplo, inconsistencias entre valor y peso, o mercancías asociadas a restricciones, pueden incrementar el control. Estas hipótesis se evalúan empíricamente con modelos interpretables y análisis de importancia de variables (SENAE, s. f.; He & Garcia, 2009).

### 6.3. Gestión de riesgo aduanero, selectividad y perfiles

La gestión de riesgo aduanero se fundamenta en la aplicación sistemática de procedimientos de evaluación para asignar recursos de control donde el riesgo es mayor. La

WCO enfatiza la necesidad de un enfoque común para identificar, tratar y comunicar riesgos, integrando información sobre mercancías, rutas, operadores y comportamientos (WCO, s. f.-a). Desde una perspectiva operacional, documentos de referencia para aduanas resaltan que la selectividad busca distinguir operaciones de bajo riesgo de aquellas que requieren control intensivo, reduciendo inspecciones individuales y adoptando análisis basado en riesgo (OSCE, s. f.).

En manuales de formación aduanera se observa un patrón recurrente de indicadores considerados en la construcción de perfiles, tales como origen, naturaleza de la mercancía, valor declarado, consistencia documental y patrones del operador (COMESA, 2023). Estos componentes respaldan la hipótesis de que, incluso sin conocer la regla interna de SENAE, las variables observables del paquete pueden aportar señal predictiva sobre la modalidad de aforo asignada.

### 6.3.1. Ciclo de gestión de riesgo en aduanas: identificar, analizar, tratar y monitorear

Los marcos internacionales de gestión de riesgo en aduanas describen un ciclo continuo: identificar riesgos, analizarlos, evaluarlos, tratarlos (por ejemplo, seleccionando controles) y monitorear resultados para mejorar el sistema. El Risk Management Compendium de la WCO enfatiza que la gestión de riesgo no es un evento único, sino un proceso iterativo que se adapta a cambios en patrones de fraude, tendencias de comercio y capacidades institucionales (WCO, s. f.-a).

En este ciclo, el 'tratamiento' del riesgo se materializa, entre otras acciones, en la selectividad de controles: canal automático para bajo riesgo; revisión documental o física para riesgo mayor; inspecciones no intrusivas como alternativa de control eficiente; y acciones posteriores (auditorías, investigaciones) para casos complejos. Para un operador courier, este proceso se observa indirectamente mediante la etiqueta de aforo y el tiempo de despacho asociado.

Una implicación clave para ciencia de datos es que el riesgo no es observable directamente: el modelo busca aproximar un riesgo latente usando variables disponibles. Por ello, este trabajo habla de 'probabilidad de aforo' y no de 'riesgo verdadero'. La diferencia es relevante para evitar afirmaciones causales: el modelo aprende patrones históricos de selección, que pueden reflejar política, restricciones técnicas o capacidad operativa, además de riesgo de incumplimiento.

### 6.3.2. Selectividad y confidencialidad de criterios: por qué no existe una regla pública simple

En términos prácticos, muchos usuarios esperan que exista una regla pública tipo 'si el valor supera X entonces físico'. Sin embargo, la selectividad en aduanas suele ser deliberadamente opaca: si los criterios exactos se conocen, pueden ser explotados para evadir control. La guía del trámite de aforo físico de SENAE señala explícitamente el uso de perfiles de riesgo para asignar modalidades de aforo (SENAE, 2025a). En el plano

normativo, documentos oficiales ratifican que los criterios de selectividad pueden considerarse de conocimiento restringido, lo que explica la ausencia de umbrales detallados en publicaciones públicas (SENAE, 2013).

Desde la perspectiva metodológica, esta opacidad convierte al problema en una tarea de aprendizaje supervisado basada en resultados históricos observables, más que en reglas explícitas. El objetivo no es reconstruir el algoritmo interno de SENAE, sino estimar probabilidades útiles para planificación operativa, con conciencia de límites éticos y de uso: el modelo sirve para anticipación logística y comunicación al cliente, no para manipular o eludir controles.

#### 6.3.3. Indicadores típicos de riesgo en aduanas y su traducción a variables de datos

Materiales de formación y guías de selectividad suelen mencionar categorías de indicadores como: origen/procedencia, naturaleza de la mercancía, valor, consistencia documental, rutas, frecuencia y comportamiento histórico del operador, entre otros (OSCE, s. f.; Common Market for Eastern and Southern Africa [COMESA], 2023). En el régimen courier, no siempre se dispone de toda la historia del importador/consignatario en un dataset operacional; por ello, este estudio se enfoca en variables observables del paquete y en agregados de comportamiento cuando existan identificadores anonimizados.

Para conectar teoría con implementación, se propone mapear los indicadores típicos a variables del dataset: (i) valor declarado y moneda como señal económica; (ii) peso real y volumétrico como diagnóstico de volumen físico y consistencias; (iii) país de origen como indicador de rutas y perfiles; (iv) texto de contenido para una aproximación semántica a la naturaleza de la mercancía; y (v) variables de manifiesto/categoría de envío como contexto operativo. Esta traducción hace explícito el puente entre gestión de riesgo y ciencia de datos.

### 6.4. Modelos predictivos de riesgo y scoring probabilístico

Un modelo predictivo de riesgo puede conceptualizarse como una función que aproxima una probabilidad condicional $P(Y|X)$, donde Y representa el evento de interés (tipo de aforo) y X representa variables observables del paquete y su contexto. En aplicaciones operativas, el valor central no es únicamente la clase predicha, sino la probabilidad asociada, pues permite diseñar umbrales y políticas de alerta (Niculescu-Mizil & Caruana, 2005). Este enfoque coincide con la noción de 'score' probabilístico ampliamente utilizada en dominios regulados, donde el objetivo es apoyar decisiones con incertidumbre explícita y trazable (Guo et al., 2017).

En el caso aduanero, la propia WCO reconoce el uso creciente de analítica avanzada y aprendizaje automático para fortalecer la gestión de riesgo y focalización de controles. En su reporte sobre adopción de inteligencia artificial y/o machine learning en aduanas, se documentan usos en gestión de riesgo y segmentación, junto con requerimientos de gobernanza, datos y ética (WCO, 2025).

### 6.4.1. Predicción probabilística como aproximación operacional

En contextos de control, un 'score' es una medida resumida que ordena casos por probabilidad de un evento (por ejemplo, ser seleccionado a control intensivo). En aprendizaje supervisado, este score suele interpretarse como una probabilidad condicional P(Y=c|X=x), donde Y es la clase (tipo de aforo) y X son variables observables del paquete. Esta formulación es útil por dos razones: (i) permite priorizar casos por riesgo relativo, y (ii) facilita decisiones operativas basadas en umbrales, por ejemplo, si P (no automático) supera 0.30 entonces alertar al cliente.

Dicho esto, es importante distinguir dos objetivos distintos del presente estudio: 'clasificar' es asignar la clase más probable y 'estimar probabilidades' representa la determinación de cuán probable es cada clase. En un problema operativo como el Courier, la segunda es más valiosa porque permite gestionar incertidumbre. Por ejemplo, un caso puede tener 0.45 de documental, 0.40 de físico y 0.15 de RX: aunque la clase ganadora sea documental, la probabilidad de físico sigue siendo significativa y debe influir en la planificación.

### 6.4.2. Modelos supervisados para clasificación multiclase

Existen múltiples familias de modelos supervisados para clasificación. Este proyecto prioriza modelos clásicos y robustos por su equilibrio entre desempeño y cuánto pueden explicar del fenómeno. Entre ellos se incluyen regresión logística multinomial, árboles de decisión y ensambles (por ejemplo, random forests o gradient boosting). En términos conceptuales, la regresión logística modela directamente probabilidades mediante una función sigmoide en el caso binario y la función softmax en el caso multiclase, lo que la hace un punto de partida natural cuando el objetivo es obtener probabilidades interpretables (James, Witten, Hastie, & Tibshirani, 2021).

Los ensambles basados en árboles capturan relaciones no lineales y efectos de interacción, por ejemplo, combinaciones de origen más rango de valor más peso, que no siempre se representan bien con modelos lineales. Sin embargo, estos modelos pueden producir probabilidades sesgadas y requieren calibración para que el score sea confiable. Por ello, el enfoque del estudio integra evaluación de calibración y, si aplica, post-procesamiento (Niculescu-Mizil & Caruana, 2005).

### 6.4.3. Modelado jerárquico: división de la tarea para reducir impacto de desbalance severo

Cuando una clase domina de manera extrema, como el aforo automático en el dataset courier, un modelo multiclase directo puede dedicar 'capacidad' a optimizar la clase mayoritaria y fallar en clases minoritarias. Una estrategia práctica es usar una arquitectura jerárquica: primero decidir si el caso es automático o no automático; luego, para los no automáticos, distinguir documental, físico intrusivo y RX. Este enfoque se alinea con la lógica operativa: el primer filtro responde a la pregunta principal del cliente: ¿se retrasará la importación?; y el segundo refina el tipo de control (He & Garcia, 2009).

Desde el punto de vista de teoría de decisión, la descomposición jerárquica permite optimizar métricas específicas para el evento raro (no automático) usando curvas Precision-Recall y selección de umbral, mientras se mantiene una clasificación razonable en el subproblema condicionado. Así, el modelo se vuelve más útil operacionalmente, aunque el macro-F1 global sea moderado.

#### 6.4.4. Ciencia de datos en contextos regulatorios: trazabilidad, gobernanza y uso responsable

Aplicar aprendizaje automático en contextos regulatorios exige considerar transparencia, trazabilidad y uso responsable. En aduanas, la WCO documenta que administraciones y proyectos de 'smart customs' están explorando analítica avanzada y aprendizaje automático para fortalecer gestión de riesgo, pero subrayan retos de gobernanza, calidad de datos, seguridad y explicabilidad (WCO, 2025).

En este estudio, el modelo no pretende reemplazar decisiones de la autoridad aduanera ni inferir reglas internas, sino apoyar decisiones operativas del courier (comunicación y planificación). Aun así, se adoptan prácticas alineadas con un enfoque riguroso: evitar fuga de información, usar partición temporal, reportar métricas por clase y analizar interpretabilidad para comprender qué variables impulsan las probabilidades. Este enfoque se alinea con recomendaciones generales de interpretabilidad y evaluación cuidadosa en sistemas que influyen en decisiones humanas (Doshi-Velez & Kim, 2017).

#### 6.4.5. Evidencia internacional sobre risk scoring y aprendizaje automático en entornos regulatorios

La construcción de puntajes de riesgo (risk scoring) a partir de información observable es una práctica consolidada en dominios regulados como el crédito y los seguros, donde el objetivo no es únicamente clasificar, sino estimar una probabilidad de incumplimiento o evento adverso para apoyar decisiones operativas. En la literatura de credit scoring, Hand y Henley (1997) describen este enfoque como un problema de clasificación estadística que utiliza variables disponibles al momento de la solicitud para estimar el riesgo, concepto análogo a estimar la probabilidad de asignación a canales de control intensivo en el contexto aduanero.

En términos metodológicos, la literatura de scoring enfatiza que el desempeño de un modelo debe evaluarse con métricas alineadas a los costos reales de error, considerando que los eventos de interés suelen ser raros. Estudios de benchmarking en credit scoring muestran que distintos algoritmos pueden dominar según la estructura del dato, el nivel de desbalance y el criterio de evaluación seleccionado (Lessmann et al., 2015). En un régimen courier, esta idea se traduce en comparar modelos y reportar métricas por clase y promedios macro, evitando conclusiones basadas en accuracy cuando el canal automático es dominante.

En aduanas, existe evidencia académica específica de la utilidad de técnicas de minería de datos para detección y perfilamiento de riesgo bajo desbalance y costos asimétricos. Por

ejemplo, Zhou (2019) plantea el problema como una clasificación costo-sensible, donde el costo de un falso negativo (no detectar una declaración riesgosa) puede ser significativamente mayor que el de un falso positivo (inspeccionar un caso legítimo). Este marco conceptual respalda, a nivel teórico, la elección de umbrales operativos y la discusión explícita del trade-off entre facilitación y control, clave en la gestión de riesgo aduanero.

Adicionalmente, el uso de aprendizaje automático en contextos regulatorios demanda explicabilidad y gobernanza. En decisiones de alto impacto, la literatura advierte que depender exclusivamente de modelos tipo 'caja negra' y luego 'explicarlos' puede ser insuficiente; se recomienda privilegiar modelos interpretables o, al menos, asegurar trazabilidad y validación rigurosa del pipeline (Rudin, 2019). En este proyecto, esto se operacionaliza mediante (i) control de fuga de información, (ii) validación temporal para simular uso prospectivo y (iii) reportes transparentes de métricas por clase, complementados con análisis interpretativo de variables.

Finalmente, estos enfoques se alinean con la evolución institucional documentada en iniciativas de 'smart customs', donde la WCO reconoce el uso creciente de analítica avanzada y aprendizaje automático para apoyar la gestión de riesgo, enfatizando requisitos de calidad de datos, seguridad y uso responsable (WCO, 2025). En conjunto, la literatura internacional respalda el enfoque predictivo-probabilístico del presente estudio como una aproximación pragmática y científicamente fundamentada para mejorar la planificación operativa sin pretender inferir las reglas internas de selectividad de la autoridad aduanera.

### 6.5. Clasificación multiclase desbalanceada y métricas de evaluación

La predicción del tipo de aforo se formula como una clasificación multiclase (k=4), con fuerte desbalance de clases. En escenarios desbalanceados, la exactitud y métricas promedio ponderadas pueden ser engañosas, pues un clasificador puede obtener alta exactitud prediciendo únicamente la clase mayoritaria. He y Garcia (2009) proponen considerar métricas sensibles a minorías, y utilizar estrategias como ponderación de clases, muestreo y evaluación por clase para reflejar adecuadamente el desempeño en eventos raros.

En particular, para eventos minoritarios, la literatura resalta la utilidad de curvas Precision-Recall (PR) por sobre Receiver Operating Characteristic (ROC) en contextos desbalanceados, dado que ROC puede presentar una interpretación visual optimista. Saito y Rehmsmeier (2015) muestran que PR es más informativa para evaluar clasificadores cuando las clases positivas son raras. Por ello, en este estudio se reporta macro-F1, balanced accuracy y PR-AUC para tareas binarias derivadas, por ejemplo, 'no automático' vs 'automático'.

#### 6.5.1. Matriz de confusión

En clasificación, la matriz de confusión resume cuántas instancias de cada clase real fueron clasificadas en cada clase predicha. En otras palabras, puede pensarse como una tabla donde las filas representan la verdad (aforo real) y las columnas la predicción del modelo.

Cada celda indica conteos. Esta tabla permite observar patrones de error relevantes: por ejemplo, si el modelo confunde con frecuencia RX con físico intrusivo, o si clasifica muchos casos documentales como automáticos.

En escenarios multiclase y desbalanceados, la matriz de confusión es más informativa que una sola métrica agregada. Sokolova y Lapalme (2009) argumentan que diferentes medidas responden de forma distinta a cambios en la matriz de confusión, y que elegir una métrica sin comprender el objetivo puede conducir a conclusiones erróneas. Por ello, este proyecto reporta métricas por clase y promedios macro (Sokolova & Lapalme, 2009).

### 6.5.2. Precisión, exhaustividad (recall) y F1

En problemas de detección de eventos raros, por ejemplo, 'no automático', las métricas clave son precisión y exhaustividad o cobertura (también llamado "recall"). La precisión responde: de todos los casos que el modelo marcó como 'no automático', ¿qué fracción realmente lo era? El recall responde: de todos los casos realmente 'no automáticos', ¿qué fracción detectó el modelo?

En un contexto courier, ambos importan, pero su prioridad depende del uso. Si el objetivo es 'alertar' al cliente sobre posibles retrasos, suele preferirse alto recall para no dejar pasar casos que sí tendrán control, aunque eso implique algunas falsas alarmas (menor precisión). Esta elección es un ejemplo de decisión de umbral y costo relativo de errores, que se formaliza en la sección más adelante.

El F1-score combina precisión y recall mediante su media. Su ventaja es ofrecer una medida única cuando se busca balance entre ambos, pero su interpretación debe hacerse por clase y no solo como promedio ponderado por frecuencia, para no ocultar desempeño pobre en clases minoritarias (He & Garcia, 2009).

### 6.5.3. Promedios macro, micro y ponderado

Cuando las clases están desbalanceadas, una alta exactitud puede ser trivial: si 96% de los casos son automáticos, un clasificador que siempre predice 'automático' logra 96% de accuracy sin aportar valor. Por ello, este estudio adopta métricas macro, que promedian desempeño de todas las clases por igual, y métricas específicas para el evento raro 'no automático' (He & Garcia, 2009).

En evaluación multiclase, se usan diferentes promedios: (i) micro-promedio agrega todos los aciertos/errores globalmente y puede estar dominado por la clase mayoritaria; (ii) macro-promedio promedia métricas por clase y es sensible a clases raras; (iii) ponderado promedia por soporte (frecuencia), lo cual se parece al micro en desbalance severo. Para un objetivo de tesis que busca comprender y predecir canales minoritarios (documental, físico, RX), el macro-F1 y reportes por clase son más informativos.

#### 6.5.4. Curvas Precision-Recall y selección de umbral para alertas operativas

Una ventaja de producir probabilidades es que la decisión final puede ajustarse con un umbral. Por ejemplo, se puede alertar al cliente si P (no automático) >= 0.30. Cambiar el umbral modifica precisión y recall: umbrales bajos aumentan recall (más alertas) pero reducen precisión (más falsos positivos); umbrales altos aumentan precisión (alertas más 'seguras') pero reducen recall (más casos reales no detectados).

Para problemas con eventos raros, Saito y Rehmsmeier (2015) muestran que las curvas Precision-Recall (PR) son más informativas que ROC, porque reflejan directamente la proporción de verdaderos positivos dentro de las predicciones positivas, que es la pregunta operativa cuando el evento es poco frecuente (Saito & Rehmsmeier, 2015). Por ello, en este proyecto se reporta PR-AUC para el evento 'no automático' y se selecciona un umbral que optimiza una función objetivo (por ejemplo, F1 o recall objetivo).

#### 6.5.5. Evaluación multiclase: enfoque One-vs-Rest y métricas por clase

Para cuatro clases (automático, documental, físico intrusivo y RX), algunas métricas se calculan por clase usando un enfoque One-vs-Rest (OvR): se considera una clase como 'positiva' y las demás como 'negativas', y se calcula precisión/recall/PR-AUC para esa clase. Esto permite evaluar de manera específica el desempeño en RX, que es la clase más rara y suele ser la más difícil.

Adicionalmente, se recomienda analizar errores cualitativamente: revisar ejemplos de alta probabilidad mal clasificados, comparar distribución de variables en falsos positivos vs verdaderos positivos, e identificar si existen patrones sistemáticos que sugieran falta de variables (por ejemplo, historial del destinatario) o ruido en la etiqueta.

### 6.6. Calibración probabilística y confiabilidad de probabilidades

La calibración evalúa si las probabilidades estimadas corresponden a frecuencias empíricas observadas. Niculescu-Mizil y Caruana (2005) demuestran que distintos algoritmos producen probabilidades con sesgos sistemáticos y que técnicas de calibración como Platt scaling e isotonic regression pueden mejorar la calidad probabilística. Guo et al. (2017) profundizan en medidas prácticas de calibración y proponen métodos simples como temperature scaling en modelos modernos, resaltando la importancia de métricas como la Expected Calibration Error (ECE). En este estudio, dado que el resultado operativo es un score probabilístico para decisión anticipada, se incluyen curvas de calibración y métricas como Brier score y ECE para evaluar confiabilidad de probabilidades (Brier, 1950; Naeini et al., 2015).

#### 6.6.1. Medición de probabilidad confiable

Una probabilidad es 'confiable' o calibrada cuando, entre todos los casos a los que el modelo asigna 0.30 de probabilidad de ser 'no automático', aproximadamente 30% efectivamente resulta no automático. Esta propiedad es crucial si se van a tomar decisiones basadas en

umbrales: sin calibración, un 0.80 podría significar realmente 0.50, llevando a sobre confianza.

Las herramientas clásicas para evaluar calibración incluyen: (i) diagramas de confiabilidad (reliability diagrams), que comparan probabilidades predichas vs frecuencias observadas por bins; (ii) Brier score, que mide el error cuadrático medio entre probabilidad y resultado binario (Brier, 1950); y (iii) Expected Calibration Error (ECE), que resume la discrepancia promedio ponderada por bins (Naeini, Cooper, & Hauskrecht, 2015). Guo et al. (2017) popularizan el uso de ECE y muestran que modelos modernos pueden ser mal calibrados, reforzando la necesidad de evaluación explícita (Guo, Pleiss, Sun, & Weinberger, 2017).

#### 6.6.2. Métodos de calibración: Platt scaling, regresión isotónica y temperature scaling

Niculescu-Mizil y Caruana (2005) comparan métodos de calibración y destacan dos técnicas ampliamente usadas: Platt scaling (ajustar una sigmoide sobre los scores) y regresión isotónica (una función monótona no paramétrica). Estas técnicas corrigen sesgos típicos: algunos modelos tienden a producir probabilidades extremas o demasiado conservadoras (Niculescu-Mizil & Caruana, 2005).

En redes neuronales, Guo et al. (2017) proponen temperature scaling, una variante de Platt scaling que ajusta un parámetro para corregir sobreconfianza. Aunque este proyecto utiliza modelos clásicos, la idea general es la misma: si el objetivo principal es reportar probabilidades, la calibración no es opcional, sino parte del rigor metodológico.

#### 6.6.3. Calibración en el caso courier

El valor aplicado del proyecto se basa en comunicar al cliente expectativas de tiempo. Para ello, se requiere que la probabilidad sea interpretable: decir '70% de probabilidad de control documental/físico/RX' tiene sentido solo si el 70% corresponde a una frecuencia aproximada en casos similares. Por esta razón, además de métricas de clasificación, este proyecto incluye métricas de calibración y curvas de confiabilidad para el evento de interés 'no automático'.

En la práctica, un score calibrado permite construir reglas de servicio: por ejemplo, (i) si P (no automático) < 0.10, prometer un tiempo estándar; (ii) si 0.10-0.30, advertir posible retraso; (iii) si > 0.30, advertir alta probabilidad de control intensivo. Estas reglas son decisiones internas del courier y deben definirse con análisis de costo/beneficio, pero la base técnica es contar con probabilidades calibradas.

### 6.7. Interpretabilidad y explicabilidad en contextos regulatorios

En contextos de control y regulación, la interpretabilidad es relevante para comprender por qué un modelo asigna alto riesgo a un caso, y para detectar posibles sesgos o fugas de información. Ribeiro, Singh y Guestrin (2016) introducen LIME como método de explicación local modelo-agnóstico, mientras que Lundberg y Lee (2017) proponen SHAP, un marco

unificado basado en valores de Shapley para atribución de importancia de variables. Molnar (2022) sistematiza métodos interpretables y sugiere buenas prácticas para comunicar resultados a usuarios no técnicos. En este proyecto, la interpretabilidad se utiliza con un doble propósito: (i) aportar contribución académica sobre factores asociados al aforo y (ii) aumentar utilidad operativa del score, evitando que sea una caja negra.

### 6.7.1. Interpretabilidad global vs explicaciones locales

Existen al menos dos niveles de explicación. La interpretabilidad global busca responder: '¿Qué variables, en general, influyen más en las predicciones del modelo?' Esto es útil para comprender factores asociados al aforo en el conjunto de datos y para orientar mejoras de captura de información.

Las explicaciones locales buscan responder: '¿Por qué este paquete en particular recibió alta probabilidad de aforo físico o RX?' Este nivel es relevante para justificar alertas a operadores internos o para auditar casos atípicos. Molnar (2022) propone una visión práctica donde ambos niveles se complementan y deben seleccionarse según el usuario y la decisión que se apoyará (Molnar, 2022).

### 6.7.2. Métodos agnósticos al modelo: LIME y SHAP

Ribeiro, Singh y Guestrin (2016) proponen LIME, un método que aproxima el comportamiento del modelo alrededor de un caso específico usando un modelo interpretable local (por ejemplo, una regresión lineal con pocas variables). Su valor radica en que funciona con cualquier clasificador, incluso si es una 'caja negra' (Ribeiro, Singh, & Guestrin, 2016).

Lundberg y Lee (2017) proponen SHAP, un marco basado en valores de Shapley para asignar contribuciones a variables, con propiedades teóricas de consistencia. SHAP permite descomponer una predicción en aportes positivos/negativos de variables, lo cual es útil para comunicar por qué un paquete fue marcado como de alto riesgo (Lundberg & Lee, 2017).

En este proyecto, aunque el prototipo se centra en desempeño y calibración, la inclusión de interpretabilidad se plantea como recomendación operativa y como contribución académica: identificar variables que consistentemente elevan la probabilidad de controles intensivos puede informar mejoras en procesos documentales y segmentación de servicio.

### 6.7.3. Interpretabilidad y gobernanza: evitar sesgos y errores por fuga de información

En contextos regulatorios o de alto impacto, la interpretabilidad se vincula con gobernanza: ayuda a detectar sesgos, errores de captura y variables que inducen decisiones espurias. Doshi-Velez y Kim (2017) argumentan que la interpretabilidad debe evaluarse de manera rigurosa, considerando el objetivo y el tipo de usuario, y no solo como un atributo 'deseable' (Doshi-Velez & Kim, 2017).

En un dataset Courier, un ejemplo de 'explicación útil' es detectar que una variable como NameAforador domina la predicción: eso indicaría fuga de información, porque el nombre del aforador es consecuencia del canal asignado. La interpretabilidad, por tanto, no solo sirve para explicar predicciones, sino para proteger la validez del estudio.

## 6.8. Enfoques no supervisados: clustering y perfiles latentes de riesgo

Los métodos no supervisados permiten explorar estructura latente cuando el objetivo no es únicamente predecir, sino comprender agrupaciones naturales de paquetes según sus características. En gestión de riesgo aduanero, la segmentación puede revelar perfiles de envíos que sistemáticamente enfrentan controles más intensivos, lo cual puede complementar al modelo supervisado y apoyar monitoreo de cambios en patrones (OSCE, s. f.; COMESA, 2023). En este trabajo, el clustering se justifica como mecanismo para: (a) descubrir perfiles latentes, (b) contrastar si ciertos clusters se asocian a mayor frecuencia de aforo documental o físico, y (c) generar hipótesis sobre variables operativas relevantes.

### 6.8.1. Uso de Clustering

El clustering se utiliza cuando no existen etiquetas para ciertos conceptos, o cuando se busca descubrir estructura latente en los datos. En este estudio, aunque sí existe etiqueta (AforoType), el clustering cumple un rol complementario: revelar perfiles de paquetes que comparten características similares y analizar si algunos perfiles están desproporcionadamente asociados a aforos más intensivos (documental, físico o RX).

Este enfoque evita que el clustering sea un 'añadido exploratorio' sin anclaje conceptual: su propósito es construir tipologías empíricas de envíos y contrastarlas con la selectividad observada, apoyando la discusión académica sobre factores asociados al control aduanero (Jain, 2010).

### 6.8.2. K-means

K-means es uno de los algoritmos de clustering más conocidos. Intuitivamente, busca ubicar k 'centros' y asignar cada punto al centro más cercano, de modo que los grupos tengan baja variabilidad interna. Su simplicidad lo hace útil como línea base, pero tiene supuestos fuertes: clusters aproximadamente esféricos y sensibilidad a escala. Por ello, es esencial estandarizar variables numéricas, por ejemplo, pesos y valor, antes de aplicar k-means (Jain, 2010).

En datos courier, la escala es crítica: el valor declarado puede variar en cientos, mientras que el peso puede variar en unidades pequeñas. Sin estandarización, una variable domina la distancia y el clustering pierde significado.

### 6.8.3. Clustering basado en densidad: HDBSCAN para perfiles con densidades distintas

Un problema habitual del courier es que existen grupos muy densos (envíos pequeños y comunes) y grupos raros (envíos atípicos). Los algoritmos basados en densidad, como

DBSCAN y sus extensiones, buscan clusters como regiones densas separadas por regiones de baja densidad. Campello, Moulavi y Sander (2013) proponen una alternativa jerárquica basada en densidad que permite identificar clusters a diferentes niveles y manejar ruido, lo que resulta atractivo para datos con perfiles muy diversos (Campello, Moulavi, & Sander, 2013).

En términos prácticos, usar un método basado en densidad puede ayudar a identificar perfiles minoritarios que podrían correlacionarse con RX o físico intrusivo, proporcionando un lente adicional para interpretar la distribución de aforos.

#### 6.8.4. Validación de Clustering

A diferencia de la clasificación supervisada, el clustering no tiene una 'verdad' única. Por ello, la validación se basa en: (i) criterios internos, como separación y cohesión, (ii) estabilidad, si los clusters se mantienen al perturbar datos o parámetros, y (iii) utilidad externa, es decir, si los clusters ayudan a explicar un fenómeno relevante, en este caso, diferencias en tasas de aforo.

Para este proyecto aplicado, la validación más importante es la utilidad externa: que los perfiles encontrados sean interpretables, por ejemplo, 'alto valor y bajo peso', 'alto volumétrico', 'origen específico', y que muestren diferencias estadísticas o prácticas en probabilidad de documental/físico/RX. Esto se presenta como complemento explicativo del modelo supervisado, no como sustituto.

### 6.9. Síntesis y modelo conceptual

Con base en la literatura revisada, se propone un modelo conceptual donde la asignación de aforo es un resultado observable de un proceso de gestión de riesgo. Las variables del paquete (valor, pesos, categoría, origen, tipo, cantidad y contenido declarado) y variables de contexto (fecha, subcourier, categoría de envío) actúan como señales observables. Estas señales alimentan un riesgo latente que se manifiesta en la asignación a un canal de aforo. Metodológicamente, esto se traduce en un modelo predictivo que estima probabilidades por clase, acompañado de evaluación de calibración y facilidad de interpretación.

#### 6.9.1. Modelo conceptual propuesto: riesgo latente, selectividad y canal observado

El modelo conceptual de este estudio puede describirse en tres capas. Primero, existe un 'riesgo latente' asociado al envío, que integra probabilidad de incumplimiento o necesidad de control. Segundo, existe un mecanismo institucional de selectividad (perfilamiento) que transforma ese riesgo y otras restricciones (categorías, capacidades operativas, políticas) en una decisión observable: el canal de aforo. Tercero, existe un resultado operativo: tiempos de despacho y experiencia del cliente.

En notación simple, se plantea: X -> R* -> Y, donde X son variables observables del paquete (valor, pesos, origen, contenido), R* es riesgo latente no observado, y Y es el tipo de aforo observado. El aprendizaje supervisado estima directamente P(Y|X), sin observar R*. El

clustering busca aproximar estructura en X que podría asociarse a niveles distintos de R* y, por tanto, a diferentes tasas de Y.

#### 6.9.2. Puente hacia la metodología: operacionalización de variables y estrategia de evaluación

Este marco teórico justifica decisiones metodológicas centrales: (i) tratar el problema como clasificación multiclase (k=4) con un enfoque jerárquico para el evento raro; (ii) evaluar con macro-F1, métricas por clase y PR-AUC para el evento no automático; (iii) incluir evaluación de calibración para que el output sea interpretable como probabilidad; (iv) aplicar clustering para perfilar estructura latente y complementar la interpretación; y (v) usar partición temporal para simular el uso prospectivo del modelo y evitar fuga temporal.

En el capítulo metodológico se detallan transformaciones específicas de variables, selección de características, manejo de desbalance y controles de validez (exclusión de variables posteriores al aforo), alineando la implementación con los principios discutidos.

## 7. Metodología de investigación

### 7.1. Metodología de ciencia de datos utilizada: CRISP-DM

En términos generales, un proyecto de aprendizaje automático no consiste solo en entrenar un modelo, sino en seguir un proceso estructurado que va desde comprender el problema hasta desplegar y monitorear resultados. Un estándar ampliamente citado es CRISP-DM (Cross-Industry Standard Process for Data Mining), que propone seis fases iterativas: comprensión del negocio, comprensión de los datos, preparación de datos, modelado, evaluación y despliegue (Chapman et al., 2000).

En este proyecto, CRISP-DM se adapta al contexto courier de la siguiente manera: (1) comprensión del negocio: definir qué significa 'aforo' y qué decisiones operativas se quieren apoyar (alertas al cliente, planificación interna); (2) comprensión de datos: analizar estructura del dataset, distribución de clases y variables disponibles; (3) preparación: limpieza, tratamiento de nulos, codificación de categóricas y construcción de features; (4) modelado: entrenar modelos multiclase y jerárquicos; (5) evaluación: validar con partición temporal, métricas robustas y calibración; (6) despliegue: generar un script reproducible y outputs (tablas/figuras) para incorporación en reportes y toma de decisiones.

El enfoque de CRISP-DM es iterativo: hallazgos del EDA pueden llevar a redefinir variables o ajustar objetivos de evaluación. Esto es especialmente importante cuando se trabaja con datos operativos donde pueden existir variables con fuga de información o cambios de definición a lo largo del tiempo.

#### 7.1.1. Adaptación operativa de CRISP-DM al caso estudiado

En este trabajo, CRISP-DM no se empleó como una referencia genérica, sino como una guía concreta para organizar las decisiones analíticas del proyecto. La fase de comprensión del

negocio permitió traducir una necesidad operativa del Courier, que es de anticipar posibles retrasos asociados al canal de aforo, en un problema de predicción probabilística multiclase. La comprensión de los datos se materializó en la revisión de campos disponibles, la cuantificación del desbalance severo del objetivo y la identificación de variables con riesgo de fuga de información. La preparación de datos incluyó limpieza, selección de campos disponibles antes del aforo, codificación de variables categóricas y construcción de variables derivadas como razones entre valor y peso.

### 7.1.2. Fases de modelado y evaluación de adaptación a la lógica de selectividad del proceso aduanero

En lugar de un enfoque exclusivamente académico, el modelado se estructuró para responder a una necesidad operativa: primero diferenciar casos automáticos de no automáticos y, en una segunda etapa, refinar la clasificación condicionada entre documental, físico intrusivo y físico RX. La evaluación no se limitó a accuracy o precisión, sino que incorporó validación temporal, curvas Precision–Recall, ROC-AUC y calibración de probabilidades, de modo coherente con la literatura sobre eventos raros y puntuaciones de riesgo (He & Garcia, 2009; Niculescu-Mizil & Caruana, 2005; Saito & Rehmsmeier, 2015).

Ilustración 1: Metodología CRISP-DM aplicada a proyecto

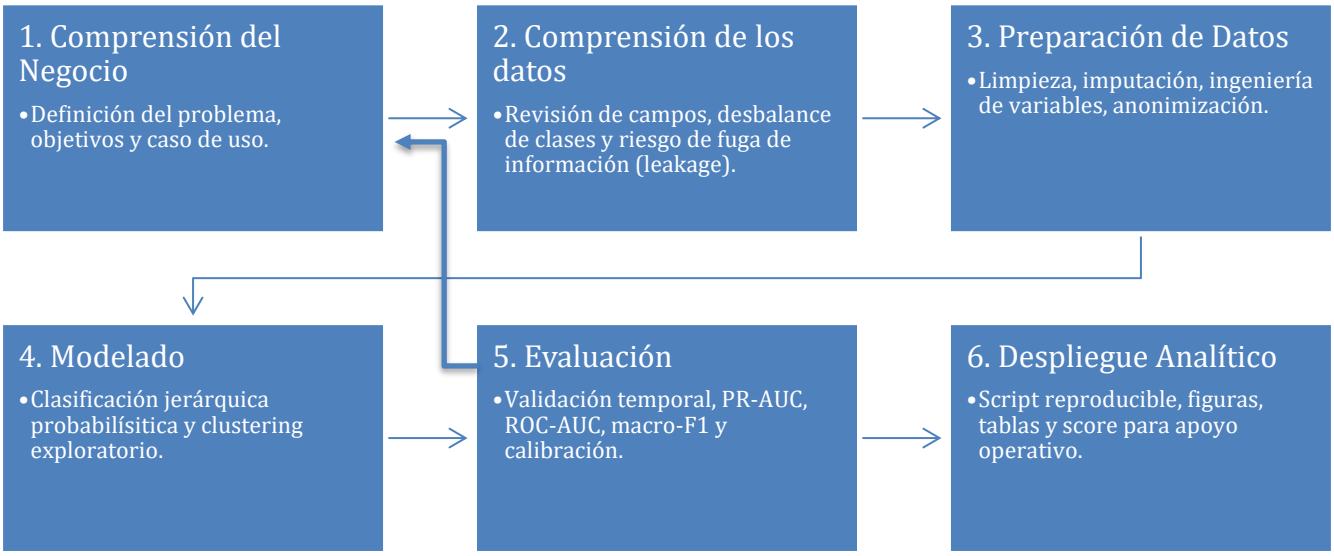



**Diagrama metodológico.** Adaptación de CRISP-DM al proyecto de predicción probabilística de aforo courier.

La Tabla siguiente resume, por fase, las decisiones metodológicas, los insumos y los resultados concretos obtenidos.

Tabla 1: Detalle de fases CRISP-DM del proyecto

| Fase CRISP-DM | Adaptación al caso courier | Evidencia / resultado |
|---|---|---|
| **Comprensión del negocio** | Definición del caso de uso: alertar probabilidad de control intensivo y apoyar la comunicación al cliente. | Problema formulado como clasificación probabilística de canales de aforo. |
| **Comprensión de los datos** | Auditoría de campos disponibles, frecuencia de clases y detección de variables posteriores al aforo. | Identificación del desbalance severo y del riesgo de leakage en NameAforador. |
| **Preparación de datos** | Limpieza, exclusión de identificadores, creación de ratios valor/peso, componentes temporales y estandarización. | Dataset analítico estructurado y variables consistentes con el momento de predicción. |
| **Modelado** | Separación del problema en etapa binaria No Automático vs Automático y etapa multiclase condicionada. | Pipeline jerárquico probabilístico con modelos lineales regularizados. |
| **Evaluación** | Partición temporal, métricas por clase, curvas PR/ROC y verificación de calibración. | Resultados operativos comparables con el escenario de uso real. |
| **Despliegue analítico** | Generación de script reproducible, tablas y figuras para el documento. | Artefactos replicables a nivel de investigación, no MLOps productivo. |

Esta explicitación por fases mejora la reproducibilidad del trabajo y permite observar que el proyecto no se restringe al entrenamiento del clasificador, sino que incluye decisiones previas de negocio, control de calidad de datos y criterios de validación alineados con el uso final del score.

### 7.1.3. Calidad de datos y control de fuga de información

En aprendizaje supervisado, la validez de resultados depende de que las variables utilizadas estén disponibles al momento de la predicción. Si se usan variables que se conocen solo después de asignado el aforo (por ejemplo, nombre del aforador), el modelo aprende una relación que no existiría en operación real. Este fenómeno se denomina fuga de información y suele producir métricas artificialmente altas.

Por ello, el proceso metodológico incluye un filtro conceptual: se priorizan variables disponibles al momento del arribo del envío o la creación del manifiesto y transmisión de la DAS. Además, se recomienda documentar explícitamente qué variables se excluyeron y por qué, como práctica de transparencia metodológica.

#### 7.1.4. Consideraciones de privacidad y anonimización en datasets operativos

Los datos courier pueden contener información personal (nombres, teléfonos, direcciones). Para fines de investigación aplicada, se han anonimizado identificadores y evitado incluir en tablas o anexos. En este estudio, el dataset se trabajó con cuidado para minimizar exposición de información personal; adicionalmente, el pipeline sugiere hash de identificadores si se incorporan variables históricas por destinatario.

Estas medidas no solo protegen privacidad, sino que mejoran la reproducibilidad académica: permiten compartir resultados agregados sin comprometer información sensible.

### 7.2. Enfoque, tipo y diseño de investigación

El estudio es de naturaleza aplicada, con componentes exploratorios y explicativos. Es **aplicada** porque busca generar una herramienta operativa para el courier. Es **exploratoria** en cuanto indaga variables del paquete asociadas al aforo, y es **explicativa** en el sentido estadístico-asociativo, sin afirmar causalidad. El diseño es cuasi-experimental computacional: se entrenan modelos con datos históricos y se evalúa su capacidad predictiva en un conjunto de prueba separado temporalmente.

### 7.3. Recolección de datos, población y muestra

La base de datos analizada corresponde a importaciones courier registradas en el período 2025-01-02 al 2026-02-20, con N=128,484 registros y 14 variables originales. La variable objetivo es AforoType con cuatro clases: Aforo Automático, Aforo Documental, Aforo Físico y Aforo Físico RX.

Tabla 2 Distribución de clases y evidencia el desbalance severo.

| Clase (AforoType) | n | % |
|---|---|---|
| **Aforo Automático** | 123.229 | 95.91 |
| **Aforo Documental** | 2.593 | 2.02 |
| **Aforo Físico** | 2.359 | 1.84 |
| **Aforo Físico RX** | 303 | 0.24 |

#### 7.3.1. Métodos de recolección de datos

Los datos utilizados en esta investigación corresponden a registros administrativos históricos generados por el sistema operativo del courier en el curso normal de su actividad. En términos metodológicos, se trata de una fuente secundaria, observacional y no experimental, ya que los eventos (envíos courier) no son intervenidos por el investigador,

sino que se analizan tal como ocurrieron en la operación real. Este enfoque es coherente con estudios aplicados en ciencia de datos orientados a la construcción de modelos predictivos a partir de datos operativos, donde el objetivo es estimar patrones y probabilidades observables y no establecer relaciones causales (Chapman et al., 2000).

La obtención de los datos se realizó mediante extracción estructurada desde la base de datos corporativa del courier. El proceso de extracción consistió en seleccionar los registros correspondientes al universo de envíos courier registrados dentro del período de análisis, incluyendo variables asociadas al paquete y al contexto operativo (por ejemplo: país de origen, pesos, valor declarado, categoría de envío, tipo de paquete, cantidad de piezas y descripción del contenido), además de la variable objetivo denominada AforoType, que representa el tipo de aforo asignado. Para garantizar trazabilidad y reproducibilidad del estudio, el dataset resultante se consolidó en un archivo tabular (formato CSV), preservando los tipos de dato relevantes y manteniendo la fecha de creación del manifiesto ManifestCreationDate como referencia temporal de cada registro.

La unidad de análisis del estudio es el envío individual (paquete) registrado en el sistema courier. La elección de esta unidad de análisis se justifica porque el objetivo del proyecto es construir una estimación probabilística por paquete que permita apoyar decisiones operativas (por ejemplo, anticipar retrasos y comunicar expectativas al cliente). En consecuencia, la estructura de los datos se organiza a nivel de registro por envío, con una etiqueta de aforo asociada y variables predictoras disponibles en el momento del arribo o del registro del manifiesto.

### 7.3.2. Criterios de inclusión/exclusión y anonimización

Para conformar la población y muestra analítica se aplicaron criterios de inclusión y exclusión orientados a preservar validez metodológica y coherencia operativa. Se incluyeron todos los envíos courier que: (a) se encontraban dentro del rango temporal definido para el estudio; (b) contenían un valor válido en la variable objetivo AforoType; y (c) contaban con información mínima necesaria en variables clave para el análisis (por ejemplo, país de origen, valor declarado o peso, según disponibilidad). Este criterio permite que el entrenamiento y evaluación reflejen el comportamiento real del proceso de selectividad, evitando sesgos derivados de registros incompletos o sin etiqueta.

Se excluyeron registros que presentaban inconsistencias críticas (por ejemplo, fechas inválidas, valores negativos o nulos no imputables en variables esenciales), duplicidades evidentes o casos donde la información disponible no representa condiciones operativas normales. Adicionalmente, se aplicó un criterio metodológico central para evitar fuga de información (data leakage): se excluyeron variables que no están disponibles al momento en que se requiere la predicción, o que son consecuencia directa del resultado (por ejemplo, campos asignados después del aforo). Esta práctica es esencial para que el desempeño reportado sea comparable con el uso real futuro del modelo y se alinea con buenas prácticas de evaluación en proyectos de minería de datos y aprendizaje automático (Chapman et al., 2000).

Respecto de privacidad y confidencialidad, el dataset fue tratado mediante un enfoque de minimización de datos: se omitieron identificadores directos de personas (nombres, teléfonos, direcciones) y códigos internos que puedan permitir reidentificación. Cuando fue necesario conservar trazabilidad técnica (por ejemplo, para agrupar por destinatario en variables históricas), se recomienda emplear identificadores anonimizados mediante funciones hash irreversibles, sin exponer datos personales en tablas, figuras o anexos. Estas medidas permiten cumplir el propósito académico sin comprometer la seguridad de la información ni la privacidad de usuarios.

### 7.4. Variables y preprocesamiento

Las variables disponibles incluyen: país de origen, subcourier, pesos (real y volumétrico), valor declarado, categoría de envío, tipo de paquete, cantidad de piezas y descripción de contenido. Se derivan variables adicionales: mes y día de semana del manifiesto; y razones como valor/peso y diferencia entre peso volumétrico y peso real. Adicionalmente, se aplica anonimización mediante exclusión de identificadores (por ejemplo, HAWB, ManifestId) en el modelado y se evita exponer valores de identificación en tablas de resultados.

Se identifica un riesgo importante de fuga de información (data leakage) en la variable NameAforador: su presencia está altamente correlacionada con aforos no automáticos, por lo que se excluye del entrenamiento para no inflar artificialmente el desempeño y para mantener validez externa del modelo.

#### 7.4.1. Linaje de datos y criterios de selección de características.

La reproducibilidad del estudio depende de dejar claro qué campos se utilizaron, qué tipo de dato representan y bajo qué criterio fueron seleccionados o excluidos. En este proyecto se privilegió un conjunto de variables disponibles antes del aforo, con capacidad explicativa, variabilidad suficiente y coherencia con la lógica operativa del régimen courier. Se descartaron identificadores directos, variables casi constantes y campos posteriores a la decisión del canal.

Tabla 3: Datos y sus características

| Campo | Tipo de dato | Rol analítico | Criterio de uso |
|---|---|---|---|
| **ManifestCreationDate** | Fecha/hora | Referencia temporal y partición | Usada para derivar mes/día y para validación temporal. |
| **AforoType** | Categórica objetivo | Variable respuesta | Define las cuatro clases del problema. |
| **OriginCountry** | Categórica nominal | Señal de procedencia | Conservada por su relevancia en perfiles de riesgo. |
| **ShippingCategory** | Categórica ordinal/nominal | Contexto del régimen | Retenida por su vínculo con límites y reglas del courier. |
| **PackageType** | Categórica nominal | Perfil físico del envío | Retenida por plausibilidad |

| | | | |
|---|---|---|---|
| | | | operativa y bajo costo de captura. |
| **DeclaredValue** | Numérica continua | Señal económica | Retenida por asociación teórica y exploratoria con control intensivo. |
| **WeightLb** | Numérica continua | Volumen/peso real | Retenida por relevancia logística y de consistencia. |
| **VolumetricWeight** | Numérica continua | Volumen estimado | Retenida por complementar al peso real. |
| **PackageQuantity** | Numérica discreta | Complejidad del envío | Retenida como proxy de consolidación. |
| **Content** | Texto libre | Naturaleza declarada | Conservada como variable susceptible de enriquecimiento semántico futuro. |

En contraste, campos como HAWB, ManifestId o NameAforador no se utilizan como predictores. Los dos primeros funcionan como identificadores administrativos sin valor generalizable; el último representa una variable posterior al proceso de selección y, por tanto, produciría un sesgo optimista al evaluar el modelo.

### 7.4.2. Técnicas de análisis: estadística, aprendizaje supervisado y no supervisado

Las técnicas de análisis empleadas integran tres niveles complementarios: análisis descriptivo, aprendizaje automático supervisado y aprendizaje no supervisado, siguiendo un enfoque de ciencia de datos estructurado mediante CRISP-DM (Chapman et al., 2000). En primer lugar, se aplicó análisis descriptivo para caracterizar la distribución de clases, cuantificar el desbalance severo y comparar estadísticos por tipo de aforo (medias, medianas y medidas de dispersión). Esta etapa es necesaria porque, en problemas desbalanceados, métricas globales como la exactitud pueden ser engañosas y ocultar un desempeño insuficiente en clases minoritarias (He & Garcia, 2009).

En segundo lugar, la modelización predictiva se abordó como un problema de clasificación multiclase (k=4) con clases altamente desbalanceadas (Automático, Documental, Físico Intrusivo y Físico RX). Dado que el objetivo operativo requiere no solo predecir una clase, sino estimar probabilidades útiles para alertas y planificación, se utilizaron modelos capaces de producir scores probabilísticos y se incorporó evaluación de calibración. La literatura señala que distintos algoritmos pueden generar probabilidades mal calibradas, por lo que se recomienda evaluarlas explícitamente y, cuando sea pertinente, aplicar técnicas de calibración para mejorar la confiabilidad del score (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017).

Debido al desbalance extremo, se priorizaron métricas robustas y orientadas a eventos raros. Para la evaluación se reportan métricas por clase y promedios macro, además de curvas Precision–Recall (PR). En escenarios con clases raras, la curva PR es más informativa que ROC para medir utilidad práctica, porque refleja la precisión real de las alertas positivas cuando el evento es infrecuente (Saito & Rehmsmeier, 2015). Asimismo, se definieron umbrales operativos en función de la relación entre precisión y recall, lo que permite al courier seleccionar el punto de operación según el costo relativo de falsos positivos (alertas innecesarias) y falsos negativos (no alertar un caso que sí enfrentará control).

Como parte del rigor metodológico, se utilizó validación temporal basada en la fecha de creación del manifiesto (ManifestCreationDate). Este diseño simula el escenario real de uso del modelo (predicción sobre casos futuros), reduce el riesgo de fuga temporal y permite evaluar generalización en presencia de cambios de distribución en el tiempo (por ejemplo, estacionalidad, cambios de demanda o ajustes normativos). Este tipo de validación es especialmente relevante cuando los datos provienen de procesos operativos que evolucionan continuamente.

En tercer lugar, se incluyó un componente no supervisado mediante clustering con el fin de identificar perfiles latentes de envíos. La justificación conceptual de este enfoque es que, aun existiendo una etiqueta observable (AforoType), el clustering puede revelar tipologías naturales de paquetes en el espacio de variables (por ejemplo, “alto valor/bajo peso”, “alto volumétrico”, “orígenes específicos”), y luego analizar si ciertos perfiles presentan tasas desproporcionadas de aforo documental, físico o RX. Este análisis complementa al enfoque supervisado aportando interpretabilidad estructural y soporte a la discusión sobre factores asociados al control selectivo.

Finalmente, para fortalecer la validez de las conclusiones, la evaluación de desempeño no se limita a una única métrica, sino que triangula evidencia mediante: (a) matriz de confusión y métricas por clase; (b) curvas PR y selección de umbral; y (c) evaluación de calibración probabilística, utilizando medidas como Brier score (Brier, 1950) y Expected Calibration Error (Naeini et al., 2015). Este conjunto de técnicas asegura que el modelo sea útil no solo como clasificador, sino como un estimador probabilístico confiable para decisiones operativas.

### 7.5. Modelos supervisados propuestos

Se evalúan dos formulaciones: (i) un modelo multiclase directo y (ii) un modelo jerárquico. El modelo jerárquico responde al desbalance severo y separa el problema en dos etapas:

Etapa 1: predicción de 'No Automático' (documental/físico/rx) vs 'Automático'.

Etapa 2: clasificación condicionada entre Documental, Físico y RX para los casos estimados como No Automático.

Esta estrategia permite definir umbrales operativos para alertas tempranas y mejorar la sensibilidad hacia clases minoritarias (He & Garcia, 2009; Saito & Rehmsmeier, 2015).

### 7.5.1. Algoritmos utilizados en la implementación final

La implementación reproducible desarrollada para este proyecto utiliza regresión logística regularizada como familia principal de clasificación. En la etapa 1 se emplea una regresión logística binaria con ponderación de clases para estimar la probabilidad de pertenecer al grupo No Automático frente a la clase Automático. En la etapa 2 se utiliza una regresión logística multinomial regularizada, restringida al subconjunto de casos no automáticos, para discriminar entre aforo documental, físico intrusivo y RX. Esta elección privilegia probabilidades nativas, estabilidad, menor riesgo de sobreajuste y mayor capacidad de trazabilidad e interpretación en comparación con modelos más complejos (James et al., 2021; Rudin, 2019).

Aunque en el marco teórico se revisan otras familias de clasificadores, por ejemplo, ensambles de árboles como random forest o gradient boosting, dichos métodos no constituyen la implementación central reportada en esta versión del estudio. La razón es metodológica y aplicada: el objetivo principal no es maximizar una métrica aislada, sino construir un score probabilístico razonablemente preciso, calibrable y defendible en un contexto regulatorio-operativo. De esta manera, el documento deja explícito que la comparación principal se realiza entre formulaciones del problema (multiclase directa vs jerárquica) y no entre una batería extensa de algoritmos heterogéneos.

### 7.5.2. Alcance del pipeline reproducible

Cuando este proyecto afirma que el trabajo es reproducible, se refiere a un pipeline analítico de investigación y no a una arquitectura completa de MLOps. El pipeline implementado cubre:

- Ingesta del archivo tabular exportado desde la operación
- Validación básica de estructura
- Preprocesamiento, construcción de variables
- Escalamiento/codificación
- Entrenamiento
- Evaluación temporal
- Calibración y exportación de tablas y figuras.

No incorpora, en esta etapa, despliegue en producción, monitoreo continuo, reentrenamiento automático ni orquestación de servicios. Esta delimitación evita sobredimensionar el alcance técnico del estudio y hace explícito que la reproducibilidad lograda corresponde al proceso de modelado y evaluación científica.

## 7.6. Evaluación, partición temporal y métricas

Para reducir fuga temporal y simular el uso prospectivo del modelo, la evaluación se realiza con partición temporal: entrenamiento con manifiestos previos a 2026-01-01 y prueba con manifiestos posteriores. Se reportan macro-F1 y balanced accuracy para multiclase, además de ROC-AUC y PR-AUC en la etapa binaria No Automático vs Automático. Como el objetivo

final es producir probabilidades, se incluye evaluación de calibración mediante curva de calibración y métricas de scoring probabilístico (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017).

### 7.6.1. Validación temporal

En datos operativos que evolucionan en el tiempo, una partición aleatoria puede mezclar casos futuros en entrenamiento y producir estimaciones optimistas. Por ello, se utiliza una validación temporal: se entrena con envíos anteriores a una fecha de corte y se prueba con envíos posteriores. Este diseño se aproxima al uso real del modelo, donde siempre se predice sobre casos futuros.

En el caso courier, factores como cambios normativos, por ejemplo, ajustes arancelarios o límites de categorías, campañas comerciales y cambios en capacidad operativa pueden alterar distribuciones. La validación temporal ayuda a detectar si el modelo generaliza a periodos más recientes y reduce la probabilidad de sobreajuste a patrones históricos.

### 7.6.2. Métricas principales

Para la clasificación multiclase (k=4) se reportan: matriz de confusión, precisión/recall por clase y macro-F1. El macro-F1 es el promedio del F1 de cada clase, por lo que penaliza fuertemente cuando el modelo ignora una clase minoritaria. Esto es importante para RX, que suele ser muy raro pero operacionalmente relevante.

Adicionalmente, para el objetivo operativo de 'alerta temprana', se construye una tarea binaria: no automático vs automático. Para esta tarea se reportan ROC-AUC y, de manera prioritaria, PR-AUC. Se elige PR-AUC porque en eventos raros la curva ROC puede dar una impresión optimista; la curva PR refleja mejor el rendimiento práctico en términos de precisión de alertas y cobertura de casos reales (Saito & Rehmsmeier, 2015).

### 7.6.3. Validación de Precisión

Precision es una métrica específica: la proporción de predicciones positivas correctas. Por ello, esta tesis distingue entre: (i) desempeño global (macro-F1, balanced accuracy), (ii) precisión como métrica por clase, y (iii) confiabilidad de probabilidades (calibración).

Validar la precisión en un contexto aplicado implica: (a) definir el caso de uso (alertar al cliente vs clasificar el tipo exacto), (b) escoger métricas alineadas al costo de errores, (c) evaluar con un esquema de partición realista (temporal), y (d) reportar no solo una cifra, sino un conjunto de evidencias: matriz de confusión, curvas PR, y análisis de umbrales. Esta 'triangulación' metodológica mejora el rigor y facilita que un lector no técnico comprenda las decisiones.

Finalmente, se recomienda complementar la validación cuantitativa con revisión cualitativa de casos: examinar ejemplos con alta probabilidad y discrepancia, identificar si existen problemas de captura de datos o variables faltantes (historial del destinatario, incidencias), y documentar límites del modelo (no causalidad, dependencia del dataset del courier).

### 7.7. Enfoque no supervisado

Se aplica clustering sobre variables numéricas y categóricas, tras codificación, para identificar perfiles latentes de paquetes. Se reporta la distribución de clases de aforo por cluster y se describen perfiles típicos. Este análisis complementa el modelo supervisado aportando interpretación estructural y soporte para hipótesis operativas.

#### 7.7.1. Especificación metodológica del clustering

El algoritmo principal seleccionado para el análisis no supervisado es K-means, aplicado sobre variables numéricas previamente estandarizadas y variables categóricas codificadas. K-means se adopta como línea base por su sencillez, interpretabilidad y facilidad de comunicar perfiles centroides en un contexto aplicado (Jain, 2010). La cantidad de clusters no se fija de manera arbitraria: se determina combinando el método del codo sobre la inercia total y el coeficiente de silueta, priorizando soluciones que ofrezcan diferenciación interpretable entre perfiles de paquetes y no solo una mejora marginal de ajuste.

Como análisis de sensibilidad, HDBSCAN se considera una alternativa útil cuando la estructura de densidades es heterogénea o cuando emergen perfiles muy pequeños y atípicos que K-means tiende a absorber en clusters mayores (Campello et al., 2013). Sin embargo, para efectos de reproducibilidad y claridad metodológica, el texto fija K-means como técnica principal del estudio y utiliza HDBSCAN únicamente como contraste conceptual para futuras extensiones del análisis.

## 8. Resultados y análisis

### 8.1. Análisis exploratorio (EDA)

El conjunto de datos muestra un predominio de aforo automático. En promedio, las clases documental y físico presentan mayor valor declarado y mayor peso volumétrico que la clase automática, mientras que RX muestra menor peso real promedio, pero un peso volumétrico relativamente mayor.

Tabla 4: Resultados del análisis exploratorio

| AforoType | n | avg_weight | avg_vw | avg_value | med_value | avg_qty |
|---|---|---|---|---|---|---|
| **Aforo Automático** | 123.229 | 7.474 | 6.849 | 139.263 | 100.0 | 1.166 |
| **Aforo Documental** | 2.593 | 9.032 | 14.41 | 227.966 | 132.91 | 1.281 |
| **Aforo Físico** | 2.359 | 12.071 | 15.476 | 274.207 | 100.0 | 1.333 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Aforo Físico RX** | 303 | 4.496 | 8.503 | 126.962 | 112.89 | 1.0 |

Ilustración 2: Distribución de clases (escala log).

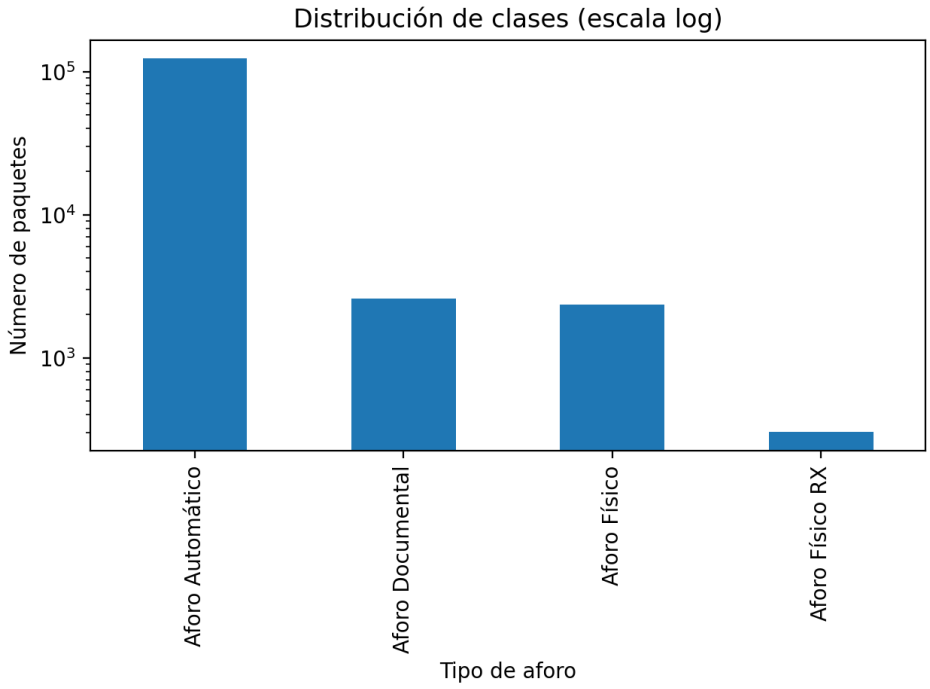



Ilustración 3: Valor declarado promedio por tipo de aforo.

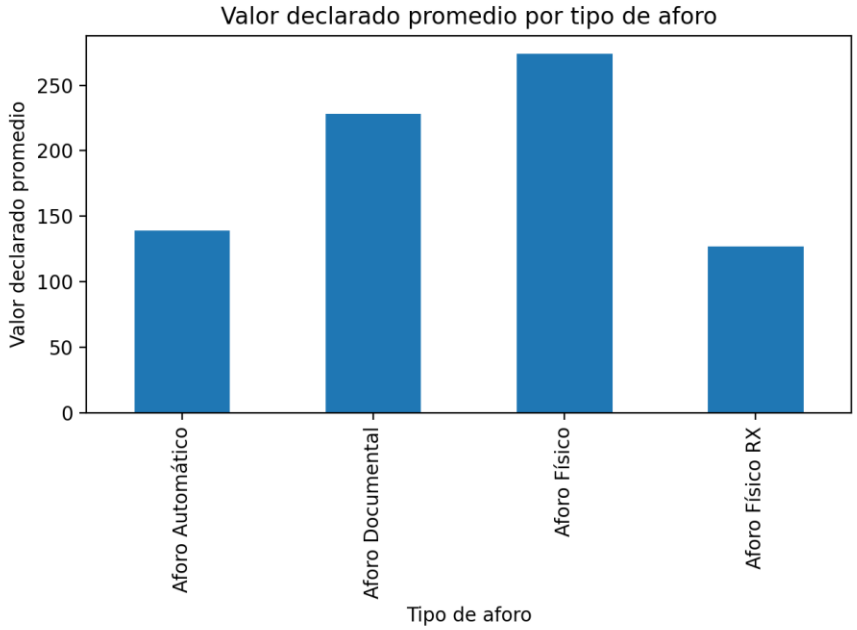


Ilustración 4: Peso real y volumétrico promedio por tipo de aforo.

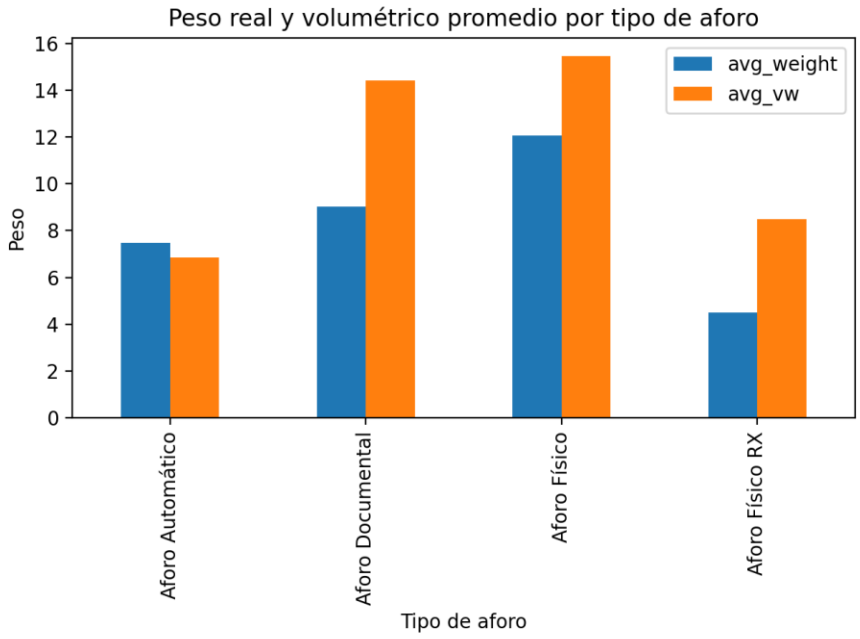



Ilustración 5: Curva Precision-Recall para detección de casos No Automático.

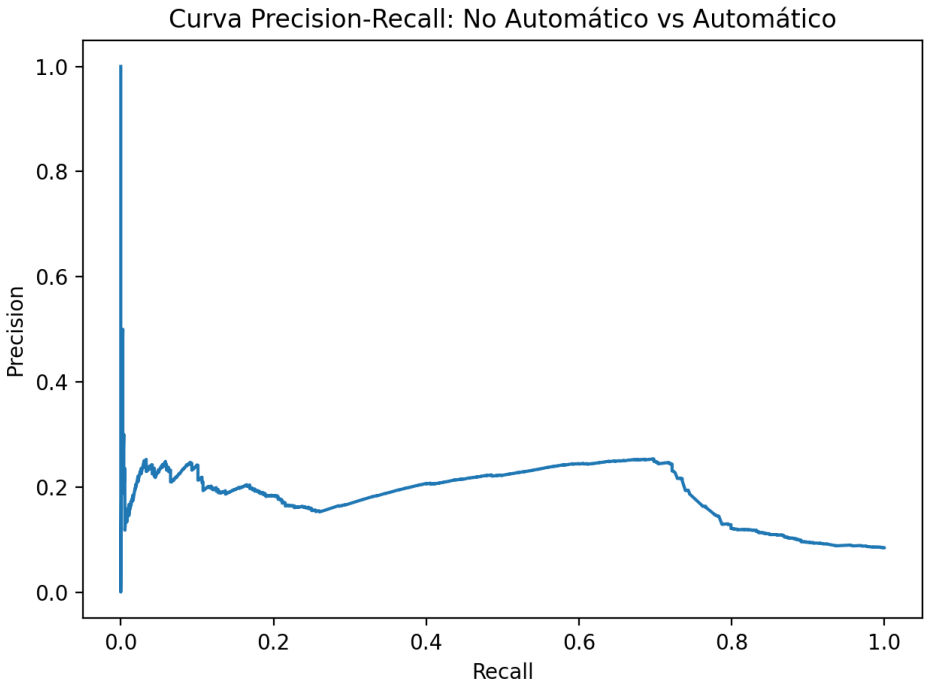
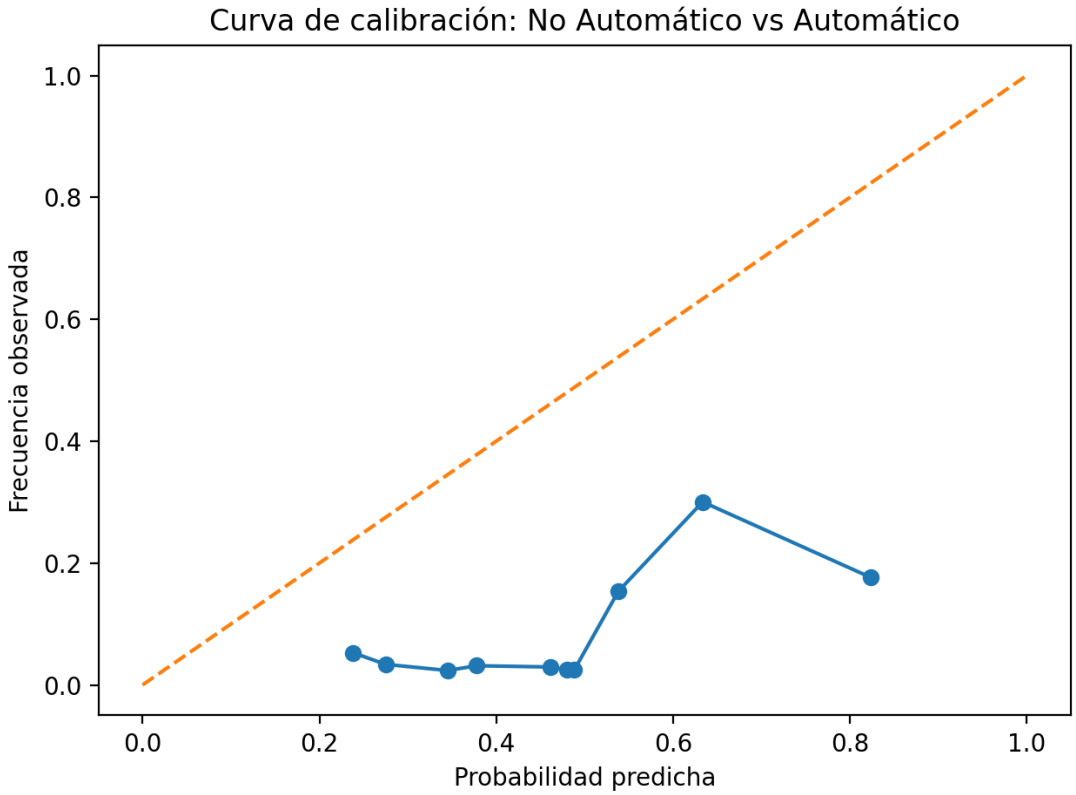


Ilustración 6: Curva de calibración para probabilidad de No Automático.




### 8.2. Desempeño del modelo jerárquico (alerta operativa)

La etapa 1 (No Automático vs Automático) alcanza ROC-AUC=0.731 y PR-AUC=0.186 en el conjunto de prueba temporal. Para un objetivo operativo de alta sensibilidad (recall≈0.80), se obtiene un umbral de probabilidad ≈0.463 con precisión≈0.121. Este resultado es consistente con la teoría de eventos raros: para detectar la minoría con alta cobertura se acepta baja precisión, lo cual es defendible cuando el costo de no alertar es alto y el costo de alerta falsa es moderado (He & Garcia, 2009; Saito & Rehmsmeier, 2015).

En la etapa 2, el desempeño multiclase condicionado presenta dificultad particular en la clase RX debido a su baja frecuencia (n=303 en el dataset total). Los resultados se reportan como parte del análisis preliminar y se discuten oportunidades de mejora mediante incorporación de variables adicionales (historial del destinatario/importador, incidencias y señales de consistencia documental) y enriquecimiento del texto de contenido declarado.

#### 8.2.1. Visualización complementaria del desempeño

A fin de responder a la observación de trazabilidad gráfica de las métricas, se incorpora explícitamente la curva ROC de la etapa binaria No Automático vs Automático. Aunque la curva Precision–Recall sigue siendo la más informativa para un evento raro, la curva ROC permite observar la discriminación general del modelo y visualizar el valor reportado de ROC-AUC=0.732.

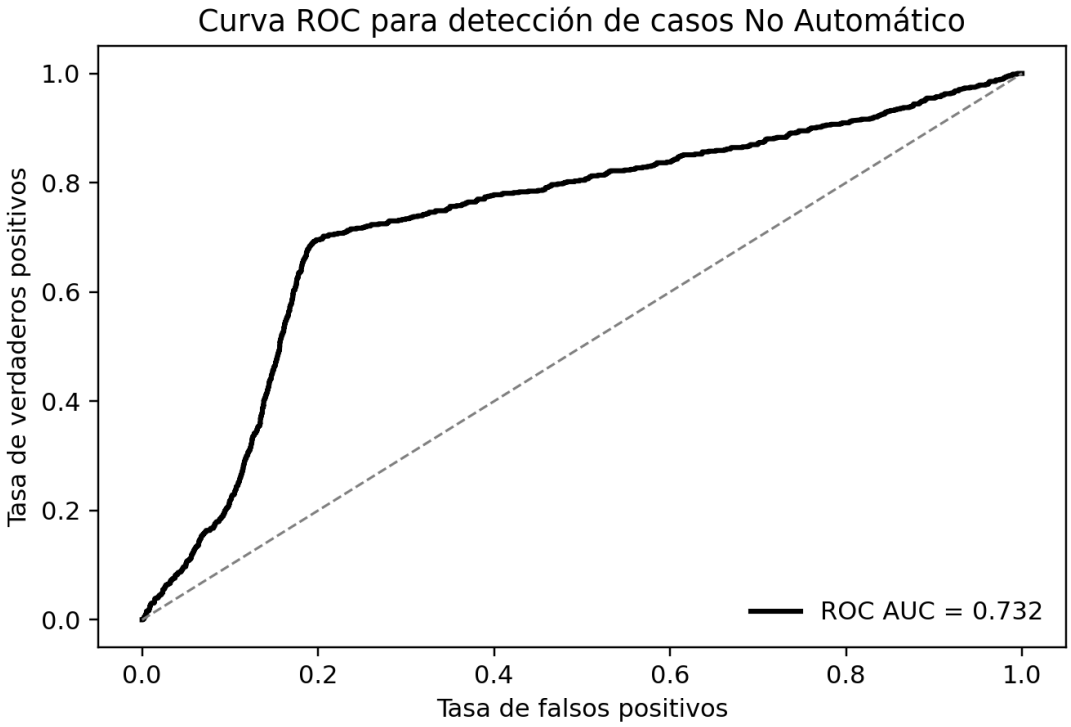



*Figura complementaria. Curva ROC para la detección de casos No Automático (etapa 1).*

### 8.2.2. Interpretabilidad global y variables con mayor influencia analítica

La elección de regresión logística regularizada facilita una lectura global de la señal predictiva, ya que el modelo asocia cambios en los predictores con incrementos o disminuciones en la probabilidad estimada. En combinación con el análisis exploratorio, el patrón más consistente ubica a DeclaredValue, WeightLb, VolumetricWeight, las razones entre valor y peso, OriginCountry, ShippingCategory y PackageType entre las variables con mayor aporte analítico para distinguir casos automáticos de no automáticos. Este hallazgo es coherente con la literatura sobre selectividad aduanera, donde valor, naturaleza de la mercancía, procedencia y consistencia físico-documental aparecen recurrentemente como indicadores de riesgo (WCO, s. f.-a; COMESA, 2023).

En términos sustantivos, los resultados sugieren que el incremento de valor declarado y de peso volumétrico tiende a concentrarse en clases sometidas a mayor control, mientras que la combinación de origen y categoría operativa ayuda a perfilar segmentos con distinta probabilidad de revisión. A la vez, el estudio identifica explícitamente que NameAforador no puede interpretarse como predictor válido, sino como una manifestación de fuga de información. Esta observación fortalece el valor metodológico del trabajo: no solo se predice un resultado, sino que se depura críticamente el conjunto de variables para evitar explicaciones no verificadas.

Tabla 5: Nivel de influencia por variable

| Variable / grupo | Nivel de influencia analítica | Interpretación operativa |
|---|---|---|
| **DeclaredValue** | Alta | Resume señal económica y posible complejidad tributaria/documental. |
| **WeightLb** | Media–alta | Aporta sobre dimensión física real del envío. |
| **VolumetricWeight** | Alta | Captura discrepancias entre volumen y peso real, relevantes para control físico/RX. |
| **Valor/peso** | Alta | Refleja consistencia económico-física del paquete. |
| **OriginCountry** | Media–alta | Introduce señal de procedencia y rutas de riesgo. |
| **ShippingCategory / PackageType** | Media | Resume reglas del régimen y tipología operativa del envío. |
| **Content (texto)** | Potencialmente alta | Variable con margen de mejora si se explota semánticamente. |

La tabla anterior debe leerse como una síntesis interpretativa y no como un ranking absoluto cerrado. Su propósito es mostrar qué grupos de variables concentran la señal más útil para el score de riesgo, y orientar futuras mejoras del pipeline, por ejemplo, la inclusión de variables históricas del destinatario o técnicas más ricas de procesamiento del contenido declarado.

### 8.3. Aportes académicos preliminares

Desde la perspectiva académica, los resultados sugieren que variables observables como valor declarado y pesos (real y volumétrico), así como categorías operativas de envío, están asociadas a la probabilidad de no ser automático. Además, la evaluación de calibración muestra que la probabilidad estimada puede interpretarse como frecuencia aproximada en rangos de score, lo cual es esencial para un uso responsable del modelo en comunicación al cliente (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017).

## 9. Discusión

Los hallazgos preliminares respaldan la viabilidad de un score probabilístico de aforo para alertas tempranas en operación courier. No obstante, la clasificación fina entre documental, físico intrusivo y RX requiere variables adicionales que capturen mejor la lógica de selectividad o la naturaleza de la carga. En particular, la clase RX es minoritaria y podría

responder a señales no contenidas en el dataset disponible, como imágenes de escaneo, inconsistencias detectadas o reglas internas específicas. En entornos regulatorios, este resultado refuerza la necesidad de alinear desempeño, interpretabilidad y gobernanza al desplegar modelos de ML (WCO, 2025; Molnar, 2022).

En primer lugar, el comportamiento observado en la etapa binaria es consistente con la literatura sobre clasificación en escenarios desbalanceados. Una PR-AUC moderada y una precisión relativamente baja a altos niveles de recall no deben interpretarse automáticamente como un mal desempeño, sino como la consecuencia esperable de buscar cobertura de una clase minoritaria dentro de un universo donde el aforo automático concentra aproximadamente el 96 % de los casos (He & Garcia, 2009; Saito & Rehmsmeier, 2015). En un uso operativo orientado a alertas, esta relación puede ser aceptable si el costo de una advertencia preventiva es menor que el costo reputacional y logístico de no anticipar un control intensivo.

En segundo lugar, la decisión de implementar regresión logística regularizada como familia principal resulta congruente con el contexto regulatorio del problema. La literatura sobre scoring y decisiones de alto impacto ha señalado que, cuando el objetivo es producir probabilidades defendibles y comprensibles, los modelos interpretables pueden ser preferibles a alternativas más opacas, aun cuando estas prometan pequeñas mejoras marginales en desempeño (Hand & Henley, 1997; Rudin, 2019). En este estudio, esa ventaja se observa en la capacidad de vincular las predicciones con grupos de variables sustantivamente plausibles, como valor, peso y procedencia.

En tercer lugar, la dificultad para separar documental, físico intrusivo y RX sugiere que la selectividad fina depende de información adicional no capturada plenamente en la base analizada. Esto no invalida el enfoque; más bien delimita su alcance: el modelo describe con utilidad la frontera entre automático y no automático, pero todavía encuentra restricciones para reproducir con mayor precisión la microdecisión interna entre modalidades intensivas. La literatura de risk scoring en aduanas y clasificación costo-sensible respalda esta lectura, ya que cuando los costos y reglas subyacentes no son observables completamente, el objetivo razonable es aproximar patrones operativos y no inferir el criterio exacto de la autoridad (Zhou, 2019).

Finalmente, el análisis no supervisado refuerza la interpretación del problema como uno de perfiles latentes más que de reglas únicas y determinísticas. La identificación de grupos relativamente homogéneos de paquetes es consistente con la idea de que la selectividad combina múltiples señales y no responde a un umbral público simple. En este sentido, el trabajo contribuye académicamente al mostrar que el modelado de aforo courier puede abordarse desde un marco de segmentación de riesgo, calibración de probabilidades e interpretabilidad, más allá de una simple predicción puntual de clases.

## 10. Limitaciones y amenazas a la validez

Limitaciones:

i. El estudio utiliza datos de un courier, por lo que la generalización a otros operadores o períodos puede requerir re-entrenamiento.
ii. La ausencia de variables del importador/destinatario limita la captura del historial de cumplimiento, reconocido como importante en perfiles de riesgo.
iii. El análisis es predictivo-asociativo; no se infiere causalidad entre variables y asignación de aforo.

Amenazas a la validez:

i. Fuga de información por variables posteriores al aforo (mitigada mediante exclusión de NameAforador);
ii. Drift temporal en perfiles de riesgo;
iii. Sesgos por concentración de origen y categorías;
iv. Desbalance severo que dificulta la evaluación de clases raras.

## 11. Conclusiones

Se concluye que es factible construir un modelo que estime probabilidades de aforo y, en particular, que detecte con utilidad operativa casos con alta probabilidad de no ser automáticos. El enfoque jerárquico permite definir umbrales que priorizan sensibilidad para alertas, coherente con evaluación en eventos raros. La separación explícita de RX como cuarta clase es conceptualmente pertinente, pero su predicción robusta requiere mayor señal explicativa (más variables y/o métodos adicionales), lo cual constituye una oportunidad de investigación aplicada futura.

Respecto del primer objetivo específico, se logró consolidar y depurar una base analítica operativa, documentando criterios de inclusión, exclusión y anonimización, así como la eliminación de variables con fuga de información. Este resultado es relevante porque transforma un conjunto de registros transaccionales en un dataset apto para análisis científico y reproduce explícitamente el linaje de datos empleado.

En relación con el segundo objetivo, la ingeniería de variables confirmó que atributos propios del paquete, valor declarado, peso real, peso volumétrico, origen y categoría de envío, contienen señal útil para aproximar la decisión de aforo. Aunque no agotan la lógica institucional de selectividad, estos factores permiten construir un score operativo con fundamento empírico y explicabilidad razonable.

Frente al tercer y cuarto objetivos, el estudio muestra que la formulación jerárquica y la evaluación con métricas robustas constituyen una respuesta metodológica adecuada al desbalance severo del problema. La utilidad principal del modelo reside en anticipar la probabilidad de no automático y no en prometer una clasificación perfecta entre todas las submodalidades intensivas. En consecuencia, la calibración de probabilidades y la selección de umbrales se consolidan como aportes centrales del trabajo para la gestión operativa.

Finalmente, respecto del quinto y sexto objetivos, el clustering y la discusión interpretativa permitieron enriquecer la comprensión académica del fenómeno. El estudio no demuestra causalidad, pero sí evidencia asociaciones consistentes y operativamente útiles entre grupos de variables y mayor probabilidad de control. En conjunto, el proyecto concluye que la predicción probabilística de aforo courier es viable, útil para la planificación y metodológicamente defendible, aunque dependiente de la disponibilidad de variables históricas y del contexto específico del courier analizado.

## 12. Recomendaciones operativas

- Usar el score de No Automático como alerta temprana para comunicación al cliente y priorización documental.
- Capturar variables adicionales de historial del destinatario/importador (frecuencia, incidencias, rectificaciones) para mejorar discriminación entre documental, físico y RX.
- Implementar monitoreo de drift: comparar distribución de variables y tasas de aforo por mes para detectar cambios en perfiles.
- Incorporar explicabilidad (SHAP/LIME) en reportes internos para justificar factores predominantes por grupo de paquetes.

## 13. Referencias


Organization for Security and Co-operation in Europe. (s. f.). Risk management and selectivity. Recuperado el 21 de febrero de 2026, de https://www.osce.org/sites/default/files/f/documents/e/f/92440.pdf

Servicio Nacional de Aduana del Ecuador. (2024b). SENAE-MEE-2-2-047-V2: Manual específico para la modalidad de despacho con canal de aforo físico no intrusivo (RX) de importación. https://www.aduana.gob.ec/gacnorm/data/2024/10/08/12/SENAE-MEE-2-2-047-V2_Manual_RX_importacion.pdf

Servicio Nacional de Aduana del Ecuador. (2025a). Realización del aforo físico de importación. https://www.gob.ec/senae/tramites/realizacion-aforo-fisico-importacion

Servicio Nacional de Aduana del Ecuador. (2025b). SENAE-MEE-2-2-004-V7: Manual específico para la modalidad de despacho con canal de aforo físico intrusivo de importación. https://www.aduana.gob.ec/gacnorm/data/2025/05/06/16/SENAE-MEE-2-2-004-V7.pdf

World Customs Organization. (s. f.-a). Risk management compendium. Recuperado el 21 de febrero de 2026, de https://www.wcoomd.org/en/Topics/Facilitation/Instrument%20and%20Tools/Tools/Risk%20Management%20Compendium

World Customs Organization. (2025). Detailed report on the adoption of artificial intelligence (AI) and machine learning (ML) in Customs. https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/facilitation/activities-and-programmes/smart-customs/public-version_detailed-report-on-the-adoption-of-ai-and-ml-in-customs.pdf

Common Market for Eastern and Southern Africa. (2023). Training manual: Risk management. https://www.comesa.int/wp-content/uploads/2023/10/Agenda-Item-5b.-TRAINING-MANUAL-RISK_MANAGEMENT_EN.pdf

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. https://www.computer.org/csdl/journal/tk/2009/09/ttk2009091263

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (ICML). https://www.cs.cornell.edu/~alexn/papers/calibration.icml05.crc.rev3.pdf

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML). https://proceedings.mlr.press/v70/guo17a.html

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3. https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In Proceedings of the AAAI Conference on Artificial Intelligence. https://ojs.aaai.org/index.php/AAAI/article/view/9602

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems. https://arxiv.org/abs/1705.07874

Molnar, C. (2022). Interpretable machine learning (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium. https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf

Comité de Comercio Exterior. (2025). Resolución No. 006-2025. https://www.produccion.gob.ec/wp-content/uploads/2025/06/Resolucion-COMEX-006-2025-signed-signed.pdf

Ministerio de Producción, Comercio Exterior e Inversiones y Pesca. (2025, 2 de junio). Gobierno Nacional aplica un arancel al régimen Courier 4x4 para frenar el uso indebido y proteger a la industria nacional. https://www.produccion.gob.ec/gobierno-nacional-aplica-un-arancel-al-regimen-courier-4x4-para-frenar-el-uso-indebido-y-proteger-a-la-industria-nacional/

Servicio Nacional de Aduana del Ecuador. (s. f.). Autorización de ingreso de mercancías mediante Mensajería Acelerada o Courier de importación. Recuperado el 21 de febrero de 2026, de https://www.gob.ec/senae/tramites/autorizacion-ingreso-mercancias-mediante-mensajeria-acelerada-courier-importacion

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Zhou, X. (2019). Data mining in customs risk detection with cost-sensitive classification. World Customs Journal, 13(2), 115–130. https://doi.org/10.55596/001c.116219

Servicio Nacional de Aduana del Ecuador. (2013). Resolución Nro. SENAE-DGN-2013-0299-RE: Disposiciones sobre confidencialidad de información empleada para perfiles de riesgo. https://www.gob.ec/sites/default/files/regulations/2025-07/Documento_Resoluci%C3%B3n-SENAE-DGN-2013-0299-RE-EMITIR-DISPOSICIONES-INFORMACI%C3%93N-RELATIVA-COMERCIO-EXTERIOR-IMPORTACIONES-EXPORTACIO.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei et al. (Eds.), Advances in Knowledge Discovery and

Data Mining (pp. 160-172). Lecture Notes in Computer Science, 7819. Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608

## 14. Anexos

1. Anexo A: Script Python