



Cuadernos de Trabajo

Gerencia de Estudios y Estadísticas Económicas
Subgerencia de Información Estadística

DETECCIÓN DE ANOMALÍAS EN LAS EXPORTACIONES ECUATORIANAS: UNA METODOLOGÍA BASADA EN EXTENDED ISOLATION FOREST

Cuaderno de Trabajo No. 150

Elaborado por:

Jair Cobos Basantes
Javier Hernández Pita
Carlos Dávila Pazmiño

Junio, 2026

BANCO CENTRAL DEL ECUADOR



Banco Central del Ecuador

Cuadernos de Trabajo

ISSN: 1390–0404

<https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Catalogo/Cuadernos/indicecuad.htm>

Cuadernos de Trabajo es una publicación que recoge los nuevos planteamientos metodológicos y los resultados preliminares de las elaboraciones estadísticas que se utilizan para los diferentes indicadores económicos realizados por el Banco Central del Ecuador.

Estos documentos están abiertos a la crítica y comentarios. En la medida en que los resultados, conclusiones y afirmaciones que contienen pueden ser objeto de cambios y enmiendas, no comprometen al Banco Central del Ecuador.

Se permite la reproducción de este documento siempre que se cite la fuente.

El análisis realizado no representa la posición del Banco Central del Ecuador o sus autoridades.

2026. © Banco Central del Ecuador

www.bce.ec



Resumen

Este estudio propone una metodología de detección de anomalías para las exportaciones ecuatorianas a partir de registros administrativos del Servicio Nacional de Aduanas del Ecuador (SENAE). Ante la ausencia de etiquetas que distingan observaciones normales y anómalas, se emplea un enfoque de aprendizaje no supervisado basado en *Extended Isolation Forest* (EIF). La metodología construye un vector de seis atributos derivados del valor unitario, que incluye *z-scores* robustos con imputación jerárquica, variaciones logarítmicas y controles estacionales. El modelo se entrena con datos de 2022–2025 y se evalúa sobre registros de enero de 2026 mediante validación por juicio experto y un ejercicio de contaminación sintética. Los resultados muestran que el EIF identifica 94 anomalías con una precisión del 81,9%. Las herramientas de inteligencia artificial explicable, como LIME y la permutación de atributos, indican que la detección se explica principalmente por desviaciones estructurales del valor unitario respecto de su trayectoria histórica y del comportamiento del mercado. En conjunto, la metodología fortalece los procesos de auditoría y control estadístico del comercio exterior.

Abstract

This study proposes an anomaly detection methodology for Ecuadorian exports based on administrative records from the National Customs Service of Ecuador (SENAE). Given the absence of labels distinguishing normal from anomalous observations, the approach relies on unsupervised learning using *Extended Isolation Forest* (EIF). The methodology constructs a six-feature vector derived from unit values, including robust *z-scores* with hierarchical imputation, logarithmic variations, and seasonal controls. The model is trained on data from 2022–2025 and evaluated on January 2026 records through expert judgment validation and a synthetic contamination exercise. The results show that EIF identifies 94 anomalies with a precision of 81.9%. Explainable artificial intelligence tools, such as LIME and feature permutation, indicate that detection is mainly driven by structural deviations in unit values relative to their historical trajectory and market behavior. Overall, the methodology strengthens audit and statistical control processes in foreign trade.

ÍNDICE

1.	Introducción	6
2.	Revisión de la literatura	7
2.1.	El valor unitario como señal de monitoreo: fundamentos económicos y estadísticos	7
2.2.	La revisión de valores unitarios como práctica de monitoreo: del proceso manual a la detección automatizada	9
2.3.	Anomalías en datos de exportación: definición, taxonomía y heterogeneidad estructural	10
2.4.	La familia Isolation Forest: principio, construcción y variantes	11
2.4.1.	Isolation Forest (iForest): fundamentos y formalización	11
2.4.2.	Construcción de un iTree	12
2.4.3.	Longitud de camino y <i>score</i> de anomalía	13
2.4.4.	Interpretación del <i>score</i> de anomalía	14
2.5.	Variantes del Isolation Forest	14
2.6.	Limitaciones del iForest clásico	16
2.7.	Extended Isolation Forest: definición formal	18
2.7.1.	Construcción de los árboles del EIF	19
2.7.2.	<i>Score</i> de anomalía del EIF	20
2.7.3.	El EIF como modelo	22
2.7.4.	Aplicaciones del Extended Isolation Forest en la detección de anomalías	22
3.	Metodología y datos	23
3.1.	Depuración de bases y consistencia de datos	24
3.2.	Creación de variables de apoyo y transformación de variables	25
3.3.	Ingeniería de atributos <i>feature engineering</i>	26
3.3.1.	Ventana de observación e imputación jerárquica	27
3.3.2.	Memoria histórica	29
3.4.	Datos	30
3.5.	Implementación de Isolation Forest	31
3.5.1.	Definición de los atributos	31
3.5.2.	Parametrización de los algoritmos	33
3.5.3.	Especificación del umbral de clasificación	34
4.	Resultados	35
4.1.	Pruebas post-estimación y explicabilidad del modelo	37
5.	Limitaciones	38
6.	Conclusiones y recomendaciones	39

Referencias	41
Anexos	44
A. Heterogeneidad del valor unitario entre categorías de exportación	44
B. Familia de algoritmos Isolation Forest	45
C. Principio de aislamiento	46
D. Impacto de la winsorización	47
E. Flujograma del tratamiento de las bases de datos	48
F. Volumen de observaciones de las bases utilizadas	49
G. Efecto de la transformación logarítmica del valor unitario	49
H. Análisis de sensibilidad del umbral	50
I. Densidad del puntaje de anomalía	51
J. Relación entre el puntaje de anomalía y el logaritmo del valor unitario	51
K. Impacto promedio de los atributos según LIME	52
L. Resultados de la permutación de atributos	52
M. Proceso de validación de resultados	52
N. Evaluación de sensibilidad mediante perturbación sintética	53

1. INTRODUCCIÓN

Todo sistema de monitoreo de comercio exterior que aspire a ser riguroso enfrenta, desde el primer momento, una tensión fundamental: la información llega en volumen, a ritmo mensual, y contiene discrepancias. En las bases de datos aduaneras que registran las transacciones de exportación, una parte de los registros refleja fielmente el funcionamiento ordinario del intercambio comercial; otra parte, en cambio, exhibe inconsistencias que pueden tener orígenes muy distintos. Algunas provienen de errores operativos —equivocaciones de digitación, unidades de medida mal declaradas, valores imputados incorrectamente— mientras que otras responden a cambios bruscos de composición en la canasta de productos exportados, a transacciones legítimas, pero estadísticamente infrecuentes, o a comportamientos que se apartan deliberadamente del patrón esperado.

La distinción entre estas categorías no es trivial, y resolverla de manera manual resultaría costosa, tardía y, sobre todo, inconsistente. La ausencia de esta herramienta introduce la vulnerabilidad de ignorar registros anómalos en las exportaciones de comercio exterior, lo que puede llegar a distorsionar la producción de las estadísticas oficiales.

Esta es la motivación central del presente trabajo: construir un sistema que detecte, de manera sistemática y reproducible, aquellas observaciones en los datos mensuales de exportaciones que se apartan del comportamiento esperado. Esta herramienta no solo optimiza los tiempos de revisión, sino que también fortalece la consistencia y eficiencia del proceso de evaluación. El objetivo no es reemplazar el juicio analítico, sino complementarlo con un mecanismo de priorización que permita identificar, con criterio cuantitativo, qué registros requieren revisión, cuáles pueden tratarse como ruido estadístico y cuáles representan señales de interés económico genuino.

Esta línea de acción responde a las recomendaciones de organismos internacionales como el Fondo Monetario Internacional, que en el manual de índices de precios de exportación e importación señala la necesidad de implementar rutinas automatizadas para detectar posibles errores y valores atípicos en los registros administrativos, con el objetivo de verificar la información y mejorar su calidad estadística (International Monetary Fund et al., 2009).

El problema, sin embargo, presenta una dificultad conceptual de fondo que lo hace más exigente de lo que podría parecer a primera vista. En los datos de exportación, la etiqueta *anomalía* no existe de forma clara ni observable. Las bases de datos aduaneras rara vez incluyen un campo que identifique, de manera confiable y consistente, si un registro es erróneo, atípico o fraudulento. En consecuencia, el problema no puede formularse como una tarea de clasificación supervisada; debe plantearse, más bien, como un problema de descubrimiento de rareza: detectar observaciones cuyo patrón multivariado se desvía significativamente de lo que ha sido históricamente frecuente, sin conocer de antemano

cuáles son esas observaciones.

Se prevé que la metodología aplicada permita capturar anomalías en los registros administrativos que no son evidentes a simple vista, debido a la complejidad de sus patrones históricos de comportamiento. En esta misma línea, se espera que los registros clasificados como anómalos por el modelo presenten una baja incidencia en enero de 2026, en concordancia con la naturaleza infrecuente de este tipo de eventos.

2. REVISIÓN DE LA LITERATURA

2.1. El valor unitario como señal de monitoreo: fundamentos económicos y estadísticos

Los datos de exportación registran múltiples dimensiones como valor FOB, peso neto, país de destino, subpartida arancelaria, entre otros, y no toda combinación de variables ofrece la misma capacidad discriminante. Para un sistema de monitoreo mensual, se requiere una señal que sea, simultáneamente, interpretable en términos económicos, sensible a inconsistencias de registro, y comparable entre distintas categorías de productos. Esa señal es el valor unitario.

En el comercio internacional las transacciones no suelen registrarse con una cotización explícita por unidad homogénea. En su lugar, las declaraciones de exportación consignan un valor monetario —típicamente el valor FOB (*Free On Board*), que excluye fletes y seguros y representa el valor de la mercancía en el punto de exportación— y una medida física de la cantidad transada. A partir de estos dos elementos se construye el valor unitario (VU) como una medida de **precio implícito**, definido formalmente como:

$$VU_{it} = \frac{FOB_{it}}{PesoNeto_{it}}, \quad \text{con } FOB_{it} > 0 \text{ y } PesoNeto_{it} > 0, \quad (1)$$

donde el subíndice i identifica el registro de exportación y t el período mensual. Esta razón cumple un rol dual. Económicamente, aproxima un precio implícito promedio, es decir, cuánto valor monetario se declara por unidad física exportada. Estadísticamente, es una variable particularmente *sensible a inconsistencias de registro*: pequeños errores en el denominador —un peso declarado en una unidad equivocada, un cero incorrecto, una magnitud fuera de escala— se traducen en saltos grandes y detectables del VU, mientras que el FOB o el peso observados de manera aislada podrían parecer “normales” por separado.

Esta sensibilidad es, lejos de ser un defecto, la principal fortaleza del VU como señal de monitoreo. En el contexto de este proyecto, el VU actúa como un amplificador de inconsistencias: tensiones entre valor declarado y cantidad física que serían difícilmente detectables mirando cada variable por separado se vuelven manifiestas cuando se las combina en esta razón. La literatura especializada ha reconocido esta propiedad desde

hace décadas. Baffes y Ajwad (2001) destacan la capacidad del valor unitario para revelar comportamientos de precio implícito en datos administrativos de comercio, mientras que Silver (2007) ofrece una discusión rigurosa sobre sus propiedades estadísticas en el contexto de índices de precios de exportación.

Sin embargo, la misma literatura advierte que el valor unitario no es un precio "puro". Cuando una categoría arancelaria agrupa variedades con características distintas —calidades, presentaciones, mezclas de productos— el VU incorpora también efectos de composición (*mix effects*). Si dentro de una misma partida arancelaria se desplaza el volumen exportado hacia variedades de mayor valor, el VU del agregado sube incluso si los precios de cada variedad permanecen estables. Formalmente, si la categoría agrega K variedades con precios p_{kt} y cantidades exportadas q_{kt} , el valor unitario del agregado es:

$$VU_t = \frac{\sum_k p_{kt} \cdot q_{kt}}{\sum_k q_{kt}}. \quad (2)$$

El VU es, por tanto, una media ponderada de precios donde las ponderaciones son las cantidades exportadas. Un cambio en la composición de las q_{kt} altera el VU aunque los p_{kt} individuales permanezcan constantes (Páramo, 2019; Silver, 2007). Esta propiedad es relevante para la interpretación de las anomalías detectadas: un VU muy elevado puede corresponder a un error de registro, a un cambio de composición inusual, o a un comportamiento de precio genuinamente atípico. El sistema detecta la rareza; la interpretación de su causa requiere análisis complementario.

Por su estructura multiplicativa y por la distribución empírica habitual de los precios de exportación —asimétricas, con colas pesadas a los extremos—, una transformación logarítmica del VU resulta estadística y conceptualmente adecuada:

$$z_{it} = \log(VU_{it}). \quad (3)$$

Con esta transformación, las diferencias relativas de precio se convierten en diferencias aditivas, las relaciones multiplicativas se linealizan, y la influencia de valores extremos puramente escalares se atenúa sin eliminar la señal anómala que interesa detectar. Esta transformación es ampliamente utilizada en la literatura de precios de comercio exterior, tanto para estabilizar la varianza como para facilitar la comparabilidad entre observaciones (Hinkle & Herrou-Aramburu, 1994; Páramo, 2019; Silver, 2007). En el sistema propuesto, $\log(VU)$ constituye la señal central alrededor de la cual se construyen los vectores de características que alimentan el modelo.

2.2. La revisión de valores unitarios como práctica de monitoreo: del proceso manual a la detección automatizada

La revisión periódica de los valores unitarios de exportación no es una práctica nueva en el análisis de estadísticas de comercio exterior. En muchas oficinas de estadística y bancos centrales, incluyendo el Banco Central del Ecuador, el monitoreo de los VU se realiza de forma manual y con frecuencia mensual: los técnicos evalúan la consistencia de los valores unitarios observados en cada período, verificando que el valor unitario de cada grupo —definido por combinaciones como partida arancelaria, exportador y país de destino— mantenga una trayectoria coherente con la del conjunto al que pertenece. Sin embargo, esta revisión se efectúa principalmente sobre el corte mensual vigente, sin incorporar de forma sistemática la dinámica histórica completa de cada grupo, lo que limita la identificación de desviaciones intertemporales más sutiles.¹

Adicionalmente, el proceso de revisión se organiza a un nivel agregado —típicamente por capítulos arancelarios de dos dígitos—, lo que introduce restricciones operativas para un análisis exhaustivo a niveles más desagregados. Como resultado, aunque el procedimiento se apoya en el conocimiento experto del analista, presenta limitaciones inherentes: su cobertura es necesariamente parcial, su reproducibilidad depende del criterio individual, y su escala no crece al ritmo de los volúmenes de datos que las administraciones aduaneras modernas generan mensualmente. Estas restricciones evidencian la necesidad de enfoques más sistemáticos y escalables para la detección de anomalías en valores unitarios.

Esta necesidad ha sido reconocida en la literatura internacional de estadísticas de comercio. A nivel de repositorios multilaterales, la evidencia muestra que los errores en las declaraciones aduaneras se manifiestan con alta frecuencia a través de valores unitarios inconsistentes, particularmente como resultado de errores en el registro de cantidades físicas (United Nations Statistics Division (UNSD), 2019). En este sentido, los sistemas de validación de bases como UN Comtrade incorporan reglas operativas basadas en valores unitarios, como la comparación del valor declarado frente a rangos históricos por categoría, identificando como candidatos a revisión aquellos registros que presentan desviaciones de gran magnitud.

En el ámbito internacional, la Comisión Europea ha avanzado hacia el uso de métodos estadísticos formales para la detección de valores atípicos en datos aduaneros. En particular, el uso de herramientas robustas diseñadas para distribuciones asimétricas, como el boxplot ajustado propuesto por Hubert y Vandervieren (2008), responde a la necesidad de adaptar los criterios de detección a la naturaleza intrínsecamente sesgada y heterogénea de los datos de comercio. Estos enfoques evidencian que los métodos clásicos, diseñados bajo

¹En este contexto, “coherencia” se refiere a la estabilidad relativa del valor unitario dentro de un grupo comparable, más que a la ausencia de variación, dado que los precios implícitos pueden fluctuar por cambios en calidad, contratos o composición del comercio.

supuestos de simetría, tienden a generar tasas elevadas de falsos positivos cuando se aplican directamente sobre valores unitarios.

Otros trabajos han explorado estrategias basadas en reglas temporales o comparaciones interperíodo para la detección de outliers en datos de comercio. Si bien estos enfoques pueden ser efectivos para identificar errores de gran magnitud —por ejemplo, desviaciones de uno o dos órdenes de magnitud—, resultan menos adecuados para capturar anomalías más sutiles, que son precisamente las más frecuentes en bases de datos altamente desagregadas (Aggarwal, 2017; Chandola et al., 2009). De manera complementaria, la evidencia empírica sugiere que la efectividad de los métodos estadísticos tradicionales depende críticamente de la estructura de los datos y de la categoría de producto analizada, sin que exista un enfoque dominante aplicable de forma general.

En conjunto, estos antecedentes convergen en un diagnóstico claro: la revisión manual de valores unitarios resulta insuficiente frente a la escala y heterogeneidad de los datos modernos de comercio exterior; los métodos estadísticos clásicos presentan limitaciones sistemáticas ante la asimetría y las colas pesadas de las distribuciones; y la detección efectiva de anomalías requiere enfoques que sean robustos, comparables entre categorías y capaces de operar de manera automatizada sobre grandes volúmenes de información.

2.3. Anomalías en datos de exportación: definición, taxonomía y heterogeneidad estructural

Antes de introducir el modelo de detección, es necesario precisar qué se entiende por *anomalía* en el contexto de los datos de exportación. Chandola et al. (2009), en su revisión seminal, definen una anomalía como una observación que no se ajusta al patrón esperado del resto de los datos. No obstante, esta definición adquiere matices particulares en bases de datos aduaneras, donde la noción de “normalidad” depende simultáneamente de factores económicos, estructurales y estadísticos.

La taxonomía estándar distingue tres tipos: (Aggarwal, 2017; Chandola et al., 2009):

- **Anomalías puntuales (*point anomalies*):** un registro individual que se desvía marcadamente del resto. En exportaciones, un ejemplo sería una partida con un valor unitario cien veces superior al rango histórico de su categoría arancelaria.
- **Anomalías contextuales (*contextual anomalies*):** observaciones que son anómalas solo en un contexto específico. Un VU elevado en flores puede ser normal en temporada de exportación de variedades premium y anómalo fuera de ella.
- **Anomalías colectivas (*collective anomalies*):** conjuntos de observaciones que individualmente parecen normales, pero cuya co-ocurrencia es inusual. Por ejemplo,

varios meses consecutivos donde un mismo exportador declara VU persistentemente fuera de rango.

En el sistema propuesto, el foco principal está en las anomalías puntuales a nivel de registro mensual, aunque el diseño de los features incorpora contexto histórico que permite capturar, indirectamente, señales de tipo contextual.

La principal dificultad metodológica en los datos de exportación radica en su heterogeneidad estructural. Las bases de datos aduaneras reúnen en una misma tabla transacciones de productos con escalas de valor unitario radicalmente distintas: un kilogramo de flores frescas, de banano, de camarón o de chatarra metálica presenta distribuciones de VU muy diferentes en magnitud y variabilidad. Esto implica que la *rareza* puede medirse con un único valor de referencia global, sino en función de lo que se considera normal dentro de cada categoría comparable. El anexo A ilustra esta heterogeneidad para algunas categorías de exportación ecuatorianas.

Esta heterogeneidad tiene consecuencias directas para el diseño del sistema: los features deben normalizarse o construirse de manera que capturen la desviación relativa a la historia de cada categoría. Adicionalmente, la heterogeneidad se extiende a la distribución estadística de los datos: las colas son pesadas, los histogramas son asimétricos, y existen valores extremos legítimos que no deben confundirse con errores. El método de detección elegido debe ser robusto ante estas características distribucionales.

2.4. La familia Isolation Forest: principio, construcción y variantes

Los métodos de detección de anomalías basados en el principio de aislamiento constituyen una familia de algoritmos que comparten una idea central: en lugar de construir un modelo explícito de normalidad para luego identificar desvíos, miden con cuánta facilidad puede separarse una observación del resto del conjunto de datos mediante particiones aleatorias. Este enfoque, propuesto originalmente por Liu et al. (2012), parte de la observación de que las anomalías tienen dos propiedades que las hacen intrínsecamente fáciles de aislar: son pocas (*few*) y están en regiones de baja densidad o en los extremos del espacio de características (*different*). En contraste, las observaciones normales son numerosas y tienden a concentrarse en regiones densas, lo que requiere muchas más particiones para aislarlas individualmente. El anexo B presenta la familia completa y sus variantes principales.

2.4.1. Isolation Forest (iForest): fundamentos y formalización

El punto de partida conceptual del Isolation Forest es una observación sobre la naturaleza de las anomalías que Liu et al. (2008, 2012) formalizan mediante dos propiedades cuantitativas. La propiedad **few** establece que las anomalías son numéricamente escasas respecto al total de observaciones: en un conjunto de n registros, el número de anomalías

n_a satisface $n_a \ll n$. La propiedad **different** establece que las anomalías tienen valores de atributos que difieren significativamente de los de las observaciones normales; es decir, están ubicadas en regiones de baja densidad o en los extremos del espacio de características.

Estas dos propiedades, tomadas conjuntamente, tienen una consecuencia geométrica directa: si se construye una partición aleatoria del espacio de características, las anomalías tenderán a quedar aisladas con muy pocas divisiones, porque están solas en sus regiones. Las observaciones normales, en cambio, comparten sus regiones con muchas otras, por lo que el árbol debe continuar particionando durante más pasos antes de aislarlas. Liu et al. (2012) demuestran que este comportamiento se mantiene incluso cuando las anomalías no están en los extremos absolutos del espacio, sino simplemente en regiones de baja densidad relativa. En el contexto de datos de exportación, las partidas con valores unitarios fuera del rango histórico de su categoría son precisamente observaciones different: su $\log(VU)$ las ubica en regiones poco frecuentadas del espacio, donde los árboles aleatorios las alcanzan rápidamente.

2.4.2. Construcción de un iTTree

Sea un conjunto de datos de entrenamiento $X = \{x_1, x_2, \dots, x_n\}$, con $x_i \in \mathbb{R}^d$, donde cada x_i es un vector de d características que representa una observación. El algoritmo construye un conjunto de m árboles de aislamiento (*isolation trees* o *iTrees*), cada uno entrenado sobre una submuestra aleatoria de tamaño ψ extraída del conjunto total.

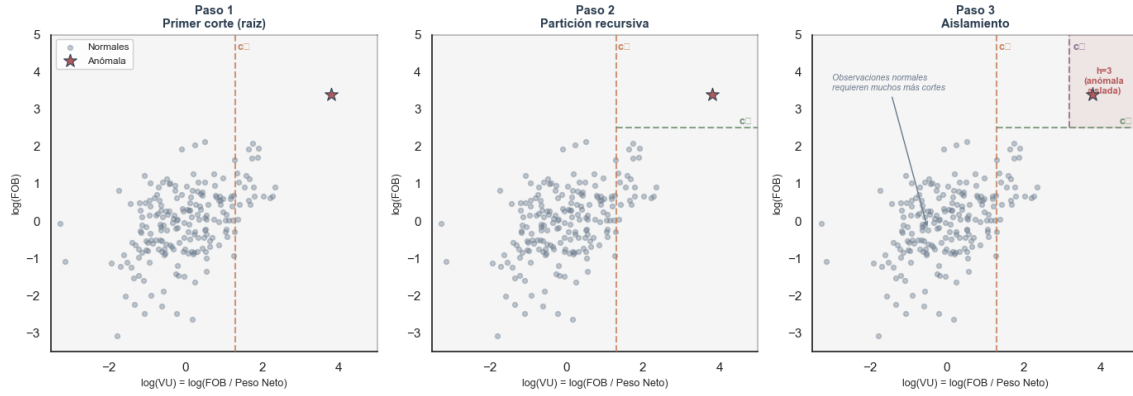
La construcción de un iTTree sigue un proceso recursivo de partición binaria. En cada nodo interno, el árbol selecciona aleatoriamente una variable $j \in \{1, \dots, d\}$ y un punto de corte c , escogido de manera uniforme dentro del rango observado de esa variable en el subconjunto de datos contenido en el nodo. Formalmente, $c \sim \text{Uniform}(\text{mín}_j, \text{máx}_j)$. El subconjunto se divide entonces en dos partes:

$$\text{Nodo izquierdo: } \{x \in S : x_j < c\} \quad (4)$$

$$\text{Nodo derecho: } \{x \in S : x_j \geq c\} \quad (5)$$

El proceso se repite recursivamente en cada rama hasta que se cumple una de dos condiciones de parada: que el nodo contenga una única observación (*aislamiento completo*) o que el árbol alcance una profundidad máxima predefinida, denotada por $l_{\text{máx}}$. Liu et al. (2012) recomiendan fijar este límite como $l_{\text{máx}} = \lceil \log_2 \psi \rceil$, donde ψ representa el tamaño de la submuestra. La Figura 1 ilustra los primeros pasos de esta construcción en el espacio $(\log VU, \log FOB)$.

Figura 1: Construcción recursiva de un árbol de aislamiento (iTree)



Nota: Construcción recursiva de un árbol de aislamiento (*iTree*) sobre el espacio bidimensional $(\log VU, \log FOB)$. Una observación anómala queda aislada en apenas tres cortes, mientras que las observaciones normales requieren más particiones.

Fuente: Elaboración propia.

2.4.3. Longitud de camino y *score* de anomalía

La longitud de camino de una observación x en un árbol T , denotada $h_T(x)$, es el número de aristas recorridas desde la raíz hasta el nodo hoja que contiene a x . Para los nodos hoja con más de una observación (cuando se alcanza la profundidad máxima antes del aislamiento completo), Liu et al. (2012) ajustan la longitud de camino añadiendo una corrección basada en la longitud esperada de un árbol de búsqueda binaria para el tamaño del subconjunto restante.

El *score* de anomalía se define promediando las longitudes de camino sobre los m árboles y normalizando por la longitud esperada $c(\psi)$:

$$E[h(x)] = \frac{1}{m} \cdot \sum_{t=1}^m h_{T_t}(x) \quad (6)$$

La función de normalización $c(\psi)$ representa la longitud de camino promedio esperada en un árbol binario aleatorio con ψ elementos. Liu et al. (2012) demuestran que esta cantidad puede aproximarse mediante:

$$c(\psi) = 2 \cdot H(\psi - 1) - \left(\frac{2 \cdot (\psi - 1)}{\psi} \right) \quad (7)$$

donde $H(k) = \sum_{i=1}^k 1/i$ es el número armónico. El *score* de anomalía normalizado adopta entonces la forma exponencial:

$$s(x, \psi) = 2^{-E[h(x)]/c(\psi)} \quad (8)$$

2.4.4. Interpretación del *score* de anomalía

El *score* $s(x) \in [0, 1]$ es una medida continua de rareza con tres rangos interpretativos bien definidos (Liu et al., 2012):

- $s(x) \approx 1$: la observación tiene una longitud de camino promedio $E[h(x)]$ muy pequeña, lo que indica que los árboles la aíslan en muy pocas particiones. Esto es consistente con las propiedades *few* y *different*: la observación está en una región poco poblada del espacio y se separa fácilmente del resto. En el contexto de exportaciones, un registro con $\log(VU)$ muy alejado del histórico de su categoría producirá un *score* cercano a 1.
- $s(x) \approx 0,5$: la longitud de camino promedio $E[h(x)]$ es aproximadamente igual a $c(\psi)$, la longitud esperada para una muestra aleatoria. La observación no puede distinguirse de lo que el modelo espera para datos normales. Este rango corresponde a registros cuyo perfil de valor unitario es indistinguible de la distribución histórica de su categoría.
- $s(x) \approx 0$: la observación requiere longitudes de camino notablemente superiores a $c(\psi)$ para quedar aislada. Esto indica que está embebida en una región densa y es difícilmente separable del resto del conjunto de datos: comportamiento claramente normal.

El *score* no afirma que una observación sea errónea, sino que cuantifica cuán inusual es su configuración de variables respecto a la historia aprendida. La Figura del anexo C ilustra la distribución de longitudes de camino y su relación con el *score* para registros de exportación.

Dos propiedades hacen que el iForest sea especialmente atractivo en términos computacionales. Primero, la construcción de los árboles se basa en submuestras de tamaño fijo ψ , lo que desacopla la complejidad de entrenamiento del tamaño total del conjunto de datos. Segundo, la evaluación del *score* para una nueva observación requiere simplemente recorrer los m árboles ya construidos, sin necesidad de recalcular distancias o densidades. En la práctica, esto se traduce en tiempos de predicción muy reducidos, lo que es relevante para un sistema de monitoreo mensual que debe procesar bases de datos de gran volumen.

2.5. Variantes del Isolation Forest

A partir del iForest original, la literatura ha propuesto extensiones orientadas a mejorar la robustez del modelo frente a distribuciones complejas y a corregir ciertas limitaciones

derivadas de su esquema de partición aleatoria (Liu et al., 2008, 2012). Estas variantes mantienen el principio de aislamiento, pero introducen modificaciones en la forma en que se construyen los cortes dentro de los árboles.

En este contexto, dos líneas relevantes de desarrollo se centran en la robustez de las particiones y en el balance en la división del espacio de datos:

- **Robust Isolation Forest (robust cut).**

El enfoque de *robust cut* surge como respuesta a la sensibilidad del iForest clásico frente a valores extremos. Como señalan Liu et al. (2012), la selección de cortes sobre el rango completo de las variables puede verse afectada por observaciones atípicas, generando particiones poco representativas.

Para mitigar este problema, las variantes robustas restringen la selección del umbral utilizando estadísticas resistentes, como cuantiles:

$$b \sim \mathcal{U}(Q_\alpha(x_j), Q_{1-\alpha}(x_j)). \quad (9)$$

Este ajuste permite que las particiones se concentren en regiones donde los datos presentan mayor estabilidad, reduciendo la influencia de valores extremos y mejorando la capacidad del modelo para identificar anomalías estructurales en lugar de simples efectos de escala (Hariri et al., 2021).

- **Fair Cut Isolation Forest.**

El enfoque de *fair cut* aborda otra limitación del iForest: la posibilidad de generar particiones altamente desbalanceadas. Como discuten Hariri et al. (2021), este tipo de divisiones puede inducir longitudes de camino artificialmente cortas, sesgando el *score* de anomalía.

Para evitarlo, se introduce un criterio de balance en la partición:

$$\left| \frac{|\{x : x_j < b\}|}{n} - \frac{1}{2} \right| \leq \epsilon, \quad (10)$$

donde ϵ controla el nivel de desbalance permitido.

Este mecanismo estabiliza la construcción de los árboles y reduce la variabilidad en la estimación de $E[h(x)]$, generando *scores* más consistentes (Bandaragoda et al., 2014).

En conjunto, estas variantes reflejan una evolución del enfoque original hacia modelos que aíslan observaciones de manera más robusta y coherente con la estructura de los datos.

En particular, la incorporación de criterios de robustez y balance anticipa la necesidad de esquemas de partición más flexibles.

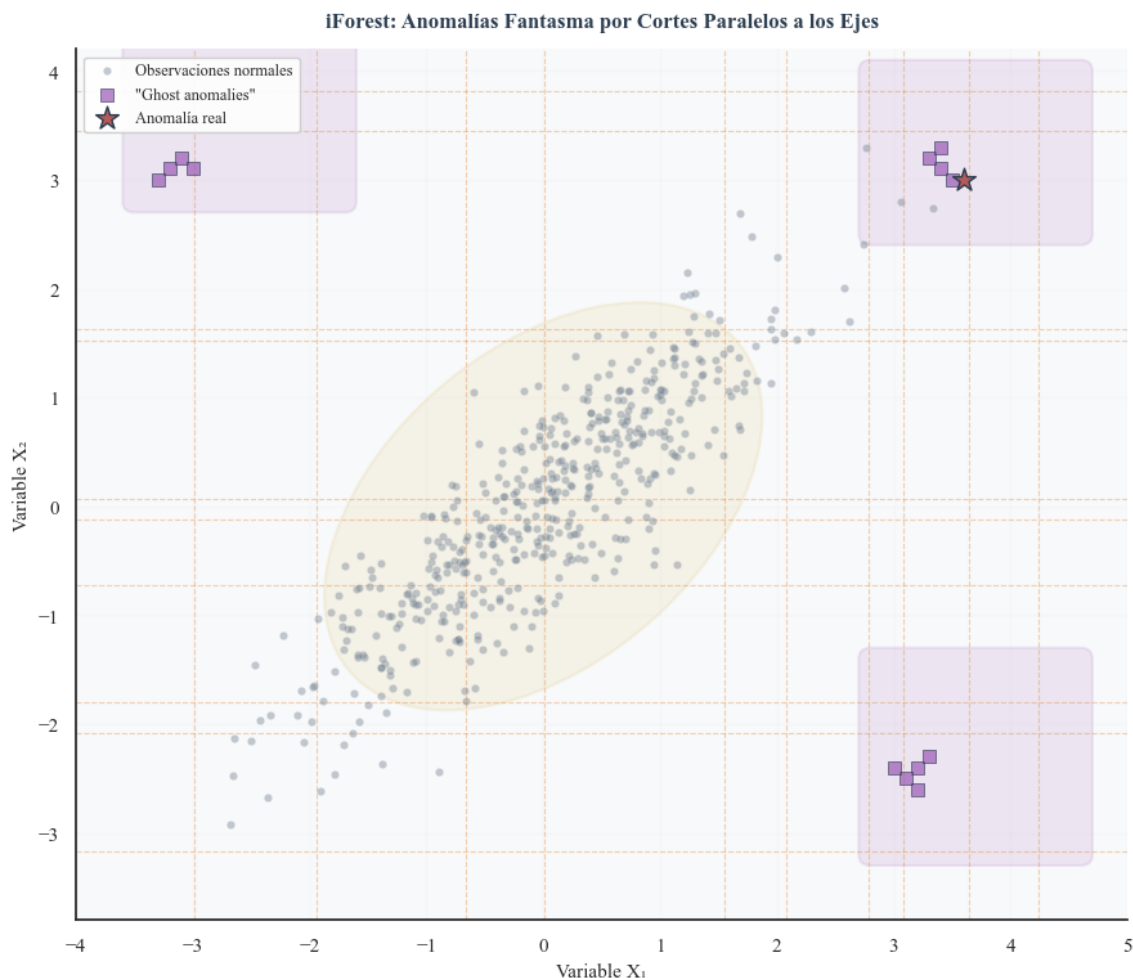
En esta línea, el Extended Isolation Forest puede interpretarse como una extensión natural de estos desarrollos, al generalizar las particiones mediante hiperplanos aleatorios, lo que permite capturar de forma más adecuada estructuras multivariadas complejas (Hariri et al., 2021).

2.6. Limitaciones del iForest clásico

A pesar de sus ventajas, el iForest, en su formulación original, presenta algunas limitaciones que han motivado el desarrollo de las extensiones descritas anteriormente y, en particular, del *Extended Isolation Forest*. En esta sección se presentan las tres más relevantes.

Limitación 1: Cortes Paralelos a los Ejes (Axis-Aligned Splits) Las particiones del iForest son siempre paralelas a los ejes coordenados: en cada nodo se selecciona una variable j y un umbral c tal que el plano de corte satisface $x_j = c$, perpendicular al eje j y paralelo a todos los demás. Hariri et al. (2021) documentaron formalmente las consecuencias de este diseño cuando los datos tienen estructura correlacional. El sesgo se manifiesta de dos maneras. Por un lado, genera las denominadas *anomalías fantasma* (*ghost anomalies*): regiones del espacio que el modelo puntúa como altamente anómalas por razones puramente geométricas —estar en las esquinas del hipercubo axis-aligned— aunque no contengan ninguna observación real ni correspondan a rareza económica genuina. Por otro lado, cuando las variables relevantes tienen correlaciones significativas, los cortes paralelos a los ejes son geoméricamente ineficientes: no pueden explotar la estructura correlacional de los datos y producen particiones que no separan bien las observaciones realmente raras de las normales. La Figura 2 ilustra visualmente ambos efectos.

Figura 2: **Limitación 1. Cortes paralelos a los ejes**



Nota. Las regiones sombreadas en púrpura son *anomalías fantasma*: esquinas del hipercono axis-aligned que iForest puntúa como altamente anómalas aunque no contengan observaciones reales ni correspondan a rareza económica. Los cortes paralelos a los ejes (líneas naranjas) no pueden explotar la estructura correlacional (elipse dorada)

Fuente: Elaboración propia.

Limitación 2: Maldición de la dimensionalidad. Un segundo problema, relacionado pero conceptualmente distinto, es la maldición de la dimensionalidad (Bellman, 1957). A medida que aumenta el número de características d , el comportamiento del espacio geométrico cambia de forma drástica: el volumen se concentra en las esquinas del hipercono, las distancias entre pares de puntos tienden a homogeneizarse y las nociones de *cercanía* y *lejanía* pierden capacidad discriminante. Zimek et al. (2012) analizan en detalle cómo este fenómeno afecta a los métodos de detección de anomalías en alta dimensión, incluidos los basados en aislamiento.

En el caso específico de iForest, la maldición de la dimensionalidad se manifiesta

del siguiente modo: cuando d es grande, disminuye la probabilidad de que un corte *axis-aligned*, seleccionado aleatoriamente sobre una variable j , separe efectivamente la observación anómala del resto. El espacio en el que emerge la rareza relevante suele depender de combinaciones de variables, no de variables individuales; por ello, los cortes que examinan una sola dimensión por vez desaprovechan justamente esa información. A medida que d crece, las longitudes de camino de las anomalías y de las observaciones normales tienden a concentrarse en valores similares, lo que comprime los puntajes hacia 0,5 y reduce la capacidad discriminante del modelo (Hariri et al., 2021).

Limitación 3: Swamping y Masking en presencia de clusters de anomalías El **swamping** ocurre cuando observaciones normales que se encuentran en la vecindad de anomalías reales reciben *scores* de anomalía artificialmente elevados. Al entrenar con el conjunto completo, las regiones de transición entre zonas densas y zonas dispersas del espacio se perciben como más anómalas de lo que realmente son, porque las anomalías cercanas “contaminan” la estimación de rareza de sus vecinas normales. El **masking** es el fenómeno inverso: cuando varias anomalías reales se agrupan en una misma región del espacio formando un pequeño cluster de rarezas, el iForest las percibe colectivamente como una zona de densidad moderada y asigna longitudes de camino más largas de lo esperado, subestimando su rareza individual. Ambos efectos se atenúan significativamente con el submuestreo descrito en la sección anterior, pero persisten en mayor grado en el iForest clásico que en el EIF, cuyas particiones oblicuas generan una separación más limpia incluso en presencia de clusters de anomalías (Hariri et al., 2021).

2.7. Extended Isolation Forest: definición formal

El Extended Isolation Forest (EIF) fue propuesto por Hariri et al. (2021) como respuesta directa a las tres limitaciones del iForest clásico descritas en la sección anterior. La motivación de los autores es precisa: las limitaciones geométricas de los cortes *axis-aligned* —las anomalías fantasma, la ineficiencia en espacios correlacionados, y la degradación con la dimensión— tienen su raíz en una restricción del mecanismo de partición, no del principio de aislamiento en sí. La solución propuesta es, por tanto, quirúrgica: reemplazar las particiones paralelas a los ejes por hiperplanos de orientación aleatoria, preservando íntegramente el principio de aislamiento y la estructura de *score*.

Hariri et al. (2021) articulan tres motivaciones formales para el desarrollo del EIF, cada una relacionada con una de las limitaciones del iForest clásico:

- **Eliminación del sesgo *axis-aligned*** Los cortes paralelos a los ejes producen mapas de *score* distorsionados que no reflejan la rareza real de los datos. Los autores demuestran analíticamente que, para distribuciones con correlaciones, los cortes *axis-aligned* asignan *scores* sistemáticamente sesgados a observaciones en las esquinas

del espacio de features, independientemente de si son realmente raras o no.

- **Corrección de las anomalías fantasma.** El EIF elimina las regiones del espacio que el iForest clásico percibe como anómalas por razones puramente geométricas. Los autores ilustran con experimentos controlados que el mapa de *score* del EIF converge al verdadero mapa de densidad inversa de los datos, mientras que el del iForest no.
- **Mejora en la robustez con la dimensionalidad** Los hiperplanos aleatorios del EIF permiten que cada partición explore combinaciones lineales de todas las variables simultáneamente, haciendo el aislamiento más robusto a medida que d crece, en comparación con los cortes que examinan una sola dimensión a la vez.

El aporte principal del EIF es, en síntesis, producir *scores* de anomalía que son más uniformes, más calibrados respecto de la rareza observada, y más estables ante la estructura correlacional de los datos —todo ello sin alterar la complejidad computacional ni la interpretabilidad del principio de aislamiento original.

2.7.1. Construcción de los árboles del EIF

En cada nodo de un árbol del EIF, la partición se define mediante un hiperplano aleatorio en \mathbb{R}^d , determinado por dos elementos:

- Un vector normal aleatorio $w \in \mathbb{R}^d$, muestreado de una distribución esférica simétrica —típicamente $N(0, I_d)$ —, que fija la orientación del hiperplano en el espacio de atributos.
- Un punto de intersección $b \in \mathbb{R}$, obtenido al proyectar las observaciones del nodo sobre la dirección w y seleccionar un valor uniforme dentro del rango de esas proyecciones. Formalmente,

$$b \sim \text{Uniform} \left(\min_{x \in S} (w^T x), \max_{x \in S} (w^T x) \right) \quad (11)$$

La regla de partición asigna cada observación del nodo a uno de dos subconjuntos según el signo de la forma lineal $w^T x - b$:

$$\text{Nodo izquierdo: } \{x \in S : w^T x < b\} \quad (12)$$

$$\text{Nodo derecho: } \{x \in S : w^T x \geq b\} \quad (13)$$

El proceso se repite recursivamente con nuevos vectores w y nuevos puntos b en cada nodo, hasta alcanzar las mismas condiciones de parada que en el iForest. La diferencia

crucial es que w es un vector en \mathbb{R}^d con orientación completamente aleatoria; por ello, cada corte puede explorar combinaciones lineales de todas las variables simultáneamente, en lugar de restringirse a un único eje.

2.7.2. *Score de anomalía del EIF*

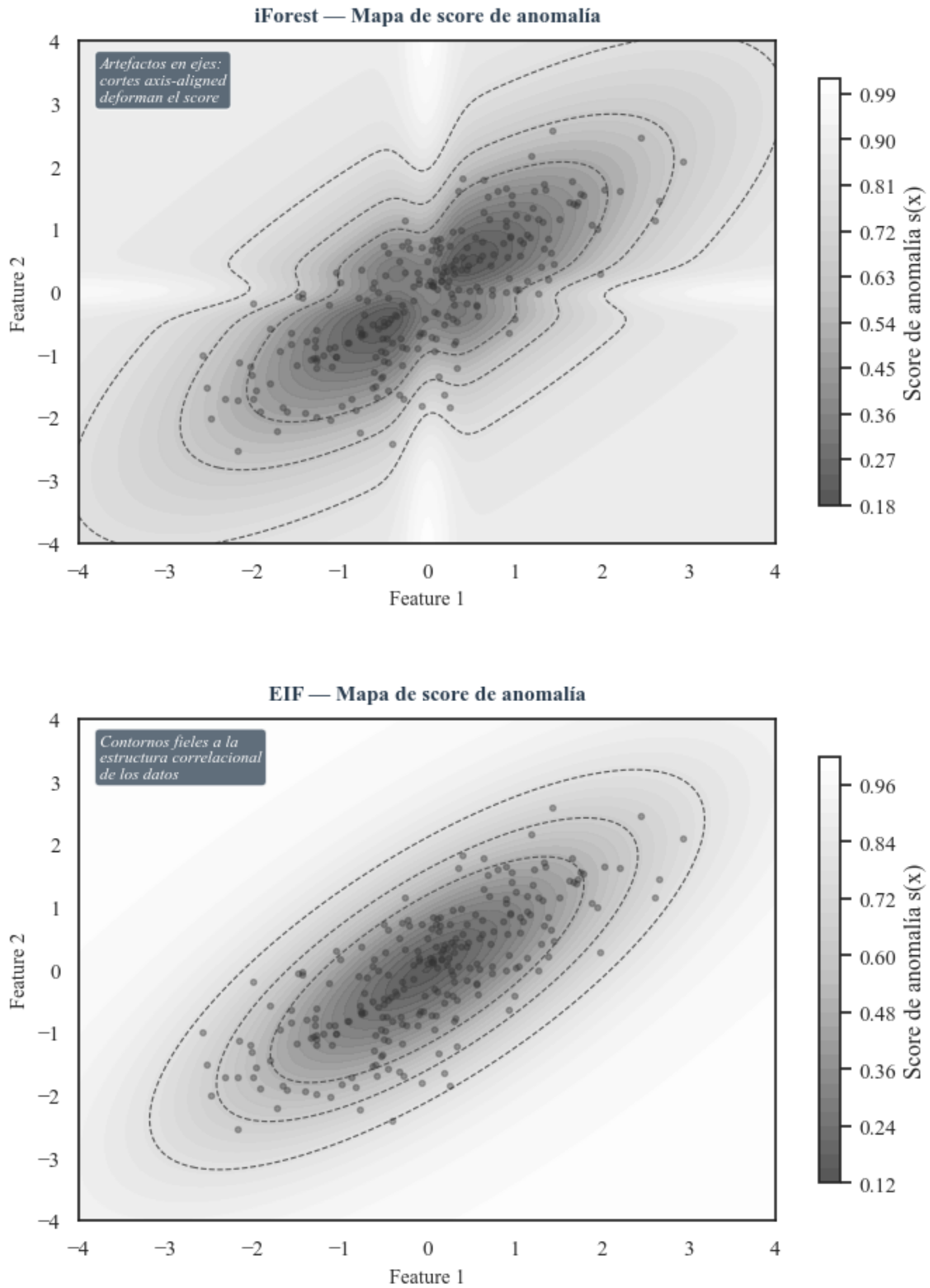
El EIF conserva íntegramente la definición del *score* del iForest original. La longitud de camino $h_T(x)$ se define de la misma manera y el *score* agregado sobre los m árboles es:

$$E[h(x)] = \frac{1}{m} \cdot \sum_{t=1}^m h_{T_t}(x) \quad (14)$$

$$s(x, \psi) = 2^{-E[h(x)]/c(\psi)} \quad (15)$$

La interpretación del *score* permanece idéntica a la descrita en la sección 4.1: valores cercanos a 1 indican anomalías probables; cercanos a 0 indican comportamiento normal; en torno a 0,5 indican indistinguibilidad de una muestra aleatoria. La diferencia fundamental está en cómo se computa $h_{T_t}(x)$: cada árbol construye sus particiones con hiperplanos oblicuos, produciendo longitudes de camino que reflejan la rareza en el espacio multivariado completo, no en ejes individuales. Hariri et al. (2021) demuestran empíricamente que los contornos de *score* del EIF reflejan fielmente la estructura geométrica de rareza de los datos, mientras que los del iForest clásico presentan distorsiones sistemáticas. La Figura 3 compara ambos mapas.

Figura 3: Mapas de *score*



Nota. Mapas de *score* de anomalía. El iForest clásico (arriba) produce contornos deformados por la geometría de los ejes y regiones de *score* elevado sin observaciones reales. El EIF (abajo) recupera contornos coherentes con la estructura correlacional de los datos, sin artefactos geométricos.

Fuente: Elaboración propia.

2.7.3. El EIF como modelo

El EIF hereda las propiedades computacionales del iForest original: una complejidad de entrenamiento de $O(m \cdot \psi \cdot \log \psi)$ y una complejidad de predicción de $O(m \cdot \log \psi)$, con un costo marginal insignificante asociado a la generación de vectores normales aleatorios en cada nodo (Hariri et al., 2021). Desde la perspectiva de la naturaleza del modelo, el EIF presenta tres características que lo hacen especialmente adecuado para el sistema desarrollado en este trabajo:

- Es no supervisado: no requiere etiquetas de anomalía para entrenarse, lo que es consistente con la realidad de los datos de exportación donde tales etiquetas no existen de forma confiable.
- Produce un *score* continuo $s(x)$ en $(0, 1]$ por observación, que funciona como una medida de rareza relativa y permite priorizar revisiones por grado de rareza en lugar de imponer una clasificación binaria rígida.
- Es flexible operativamente: el *score* continuo puede convertirse en una decisión de revisión mediante la elección de un umbral τ , definido según criterios operativos que se describen en la sección metodológica del trabajo.

Con esta definición formal del EIF, el marco teórico queda completo. Los fundamentos económicos del valor unitario, la taxonomía de anomalías en datos de exportación, y la formalización de la familia Isolation Forest —incluyendo sus limitaciones y la respuesta que el EIF ofrece a cada una de ellas— proveen el sustento conceptual sobre el que se construye el sistema de detección. La operacionalización de estos conceptos en decisiones metodológicas concretas —selección de muestras, construcción de features, elección del umbral— se desarrolla en la sección metodológica del trabajo.

2.7.4. Aplicaciones del Extended Isolation Forest en la detección de anomalías

La familia Isolation Forest, y en particular su extensión EIF, ha sido aplicada en una amplia variedad de dominios para la detección de anomalías, lo que evidencia su versatilidad y robustez frente a distintas estructuras de datos.

En el ámbito de la ciberseguridad Moomtaheen et al. (2024) evalúan el desempeño del EIF para la detección de intrusiones en redes, utilizando datos de tráfico generados con la herramienta *Zeek*. El estudio demuestra que el EIF supera al Isolation Forest clásico en la identificación de patrones anómalos en entornos de alta dimensionalidad, donde las correlaciones entre variables son relevantes y los cortes paralelos a los ejes producen sesgos sistemáticos. Los autores destacan la capacidad del EIF para operar sin etiquetas previas,

lo que resulta especialmente valioso cuando la naturaleza de la anomalía no se conoce de antemano, condición que también caracteriza los registros administrativos de exportación.

En la detección de fraude financiero, Rajeev y Devi (2022) aplican Isolation Forest para la identificación de transacciones fraudulentas con tarjeta de crédito, un dominio caracterizado por un marcado desbalance entre observaciones normales y anómalas, análogo al que se presenta en los registros aduaneros. Sus resultados —junto con los obtenidos por la familia EIF en contextos similares— confirman que el principio de aislamiento es particularmente eficaz cuando las anomalías son escasas y estructuralmente distintas del comportamiento normal, propiedades que también se verifican en los datos de comercio exterior.

En una línea similar, Shao et al. (2025) desarrollan una metodología basada en EIF con umbrales adaptativos para el monitoreo de la salud estructural de infraestructuras, proponiendo un mecanismo de ajuste dinámico del umbral de clasificación en función de las características estadísticas de la distribución de puntajes. Este enfoque es especialmente relevante para el presente trabajo, que también enfrenta el desafío de definir un umbral de clasificación robusto sobre una distribución de puntajes heterogénea.

Finalmente, Ayoub et al. (2022) proponen una combinación de EIF con el algoritmo de agrupamiento *Fuzzy C-Means* para entornos donde los datos presentan fronteras difusas entre clases, aprovechando la capacidad del EIF para asignar puntajes de rareza individuales junto con la del agrupamiento difuso para capturar estructuras de densidad más complejas. Este enfoque ilustra la flexibilidad del EIF como componente dentro de arquitecturas metodológicas más amplias.

En conjunto, la evidencia empírica revisada muestra que el EIF ha demostrado un desempeño consistentemente superior al Isolation Forest clásico en contextos de alta dimensionalidad, estructuras correlacionales complejas y distribuciones asimétricas. Estos son precisamente los rasgos que caracterizan los datos de exportación ecuatorianos analizados en este trabajo, lo que refuerza la pertinencia de seleccionar el EIF como algoritmo principal para el sistema de detección propuesto.

3. METODOLOGÍA Y DATOS

La presente investigación propone un marco metodológico basado en aprendizaje automático (*machine learning*) para la detección de anomalías en los registros administrativos de exportación. Dado que no existe una etiqueta observada que permita clasificar los datos de manera previa como anómalos o no, y considerando, además, la complejidad inherente a la dinámica de las transacciones comerciales, el problema se abordó desde un enfoque no supervisado. En este contexto, se seleccionó el algoritmo *Isolation Forest* por su buen desempeño en entornos de alta dimensionalidad y por su capacidad para identificar

observaciones atípicas de manera eficiente.

3.1. Depuración de bases y consistencia de datos

En la fase inicial de la metodología, se depuraron las bases de datos de los registros aduaneros provenientes del Servicio Nacional de Aduanas del Ecuador (SENAE), con el fin de garantizar la consistencia de la información y evitar errores que introduzcan ruido estadístico en el modelo.

Para este ejercicio se utilizaron dos bases de datos: una base histórica, que contiene las observaciones del período 2022 - 2025, y una segunda base con información de enero de 2026, sobre la cual se aplica la detección de anomalías. Esta distinción es importante ya que el tratamiento de la base histórica difiere, en algunos aspectos, del aplicado a la base de enero de 2026.

Estructuralmente, la **base histórica** se define como el conjunto de entrenamiento (*train set*), del cual se extraen los parámetros estadísticos, las medianas y los umbrales de normalidad. Por su parte, la **base de enero de 2026** se establece como el conjunto de prueba (*test set*). Esta estructuración evita la fuga de datos (*data leakage* (Miljković, 2010)); es decir, impide que la información contenida en el *test set* contamine las operaciones de *feature engineering* calculadas exclusivamente a partir del *training set*. De esta manera, el modelo se enfrenta a datos no observados previamente y los contrasta con la estructura de normalidad aprendida durante el entrenamiento.

Las bases de datos provenientes de la SENAE se encuentran estructuradas en su nivel mínimo de desagregación, es decir, cada observación corresponde a una declaración aduanera independiente, en la que es posible observar el ítem transaccional. No obstante, para facilitar el análisis estadístico, estos registros se agruparon por partida arancel nacional, exportador y fecha. Una vez agregada la información, se realizó una limpieza inicial de las variables numéricas relevantes (*valor FOB* y *peso neto*), que consistió en eliminar las observaciones con valores faltantes, negativos o inconsistentes.

En las variables categóricas se aplicaron procesos de limpieza y estandarización de caracteres. En particular, la variable *partida* se homologó a 10 dígitos, mientras que la variable *país* se normalizó bajo la codificación *Latin-1/ASCII*, con el objetivo de asegurar la integridad de los grupos.

Con el fin de reducir el impacto de observaciones atípicas o inconsistentes en la base de datos histórica que introducen ruido al modelo, se aplicó una técnica de winsorización en los percentiles 5 y 95. Este método consiste en reemplazar los valores atípicos que se encuentran en las colas de la distribución por los límites correspondientes a dichos percentiles. Esta técnica robustece la precisión del modelo al mitigar el efecto de los registros erróneos o eventos aislados de baja probabilidad. De esta forma, el algoritmo se concentra en detectar anomalías reales y disminuye la tasa de falsos positivos. El impacto

de la winsorización se puede observar en el anexo D.

$$X_w = \begin{cases} P_5 & \text{si } X < P_5 \\ X & \text{si } P_5 \leq X \leq P_{95} \\ P_{95} & \text{si } X > P_{95} \end{cases} \quad (16)$$

Cabe destacar que este proceso se aplicó únicamente a la base histórica. No se realizó este ajuste sobre los datos del mes de ejercicio (enero), ya que ello constituiría un error metodológico: implicaría suprimir anomalías de interés y neutralizar la capacidad de detección del algoritmo frente a variaciones extremas reales, introduciendo así un sesgo de subestimación.

3.2. Creación de variables de apoyo y transformación de variables

Tras la normalización de las observaciones y la limpieza de los datos, se calcularon variables auxiliares que resultaron relevantes tanto para el *feature engineering* como para la lógica de transparencia y explicabilidad del modelo. Estas fueron las siguientes:

- **Valor Unitario (VU):** Calculado como se explicó en el apartado teórico.
- **Transformaciones logarítmicas:** Se realizaron transformaciones logarítmicas del *valor FOB, peso neto y valor unitario*.

$$\ln(\text{Valor FOB}), \ln(\text{Peso Neto}), \ln(\text{Valor Unitario}) \quad (17)$$

- **Número de observaciones históricas:** Esta es una variable categórica que captura la frecuencia histórica de las observaciones, es decir las veces que un exportador realizó una operación transaccional de una partida específica a un país determinado (combinación *exportador-país-partida*).

$$N_{e,p,k} = \sum_{t=1}^T \mathbf{1}(X_{e,p,k,t} > 0) \quad (18)$$

donde:

- $N_{e,p,k}$: número de observaciones históricas para la combinación exportador–país–partida².
- $X_{e,p,k,t}$: Registro de la operación en el período t .

²Donde e denota el exportador, p el país de destino y k la partida arancelaria. Esta notación será usada en el resto del documento.

- $\mathbf{1}(\cdot)$: función indicadora, que vale 1 si hubo operación y 0 en caso contrario.
- T : total de períodos observados.

Esta variable permite monitorear la continuidad operativa del exportador. Su propósito es construir la trayectoria transaccional del operador, permitiendo al modelo distinguir entre flujos comerciales consolidados y operaciones esporádicas que podrían representar una anomalía. Para este análisis, se definió que, si un operador tenía tres o más observaciones en un año³, se considera que tiene historia a ese nivel de desagregación; si tiene menos, no la tiene. Esta aclaración es clave para el *feature engineering*, que será explicado con mayor detalle en las próximas secciones.

- **Modelado de ciclos estacionales (*sen_mes* y *cos_mes*)**. Se aplicaron componentes de Fourier mediante transformaciones trigonométricas de seno y coseno para la variable *mes*, con el objetivo de proyectar los meses en un espacio bidimensional circular y capturar la proximidad temporal entre el cierre (diciembre) y la apertura (enero) del año. Esto permite que el modelo considere los ciclos de producción y exportación para controlar el efecto de la estacionalidad.

$$\text{sen_mes} = \sin\left(\frac{2\pi \cdot \text{mes}}{12}\right), \quad \text{cos_mes} = \cos\left(\frac{2\pi \cdot \text{mes}}{12}\right) \quad (19)$$

Para una mejor comprensión de las fases iniciales de la metodología, se construyó un flujograma que resume el proceso llevado a cabo sobre las bases de datos utilizadas en el ejercicio del modelo; dicho esquema está detallado en el anexo E.

3.3. Ingeniería de atributos *feature engineering*

Más allá del procesamiento básico de los datos, la transformación de datos brutos en variables diseñadas estratégicamente es fundamental para alimentar al modelo con información relevante y de calidad. En este sentido, el *feature engineering*, también conocido como *ingeniería de atributos*, consiste en construir distintas representaciones de las variables predictoras con el fin de aumentar la eficacia del modelo (Kuhn y Johnson, 2020). Con el objetivo de afinar el modelo y mejorar sus resultados, se llevó a cabo un ejercicio riguroso de *feature engineering* sobre las variables de interés.

Es importante recordar la alta heterogeneidad intrínseca del *valor unitario* en las operaciones de comercio exterior. Esto implica que un mismo exportador puede registrar transacciones de una misma partida hacia distintos destinos; sin embargo, tanto la especificación técnica del producto como su calidad pueden variar significativamente según el

³Como el mes de enero corresponde al inicio del año 2026, no existiría historia bajo esta lógica, por lo que para este mes se amplió el campo de análisis, abarcando también el 2024.

mercado de destino. En este contexto, se adopta la premisa de consistencia operativa del exportador; es decir, se asume que la relación entre el destino de exportación, el producto y su calidad permanece estable en el tiempo. Bajo esta condición, cualquier desviación significativa del valor unitario respecto de su trayectoria esperada puede ser anomalía.

No obstante, se destaca que la relación *exportador-partida-país* presenta una intermitencia temporal en los registros aduaneros. Es decir, no siempre se observan flujos continuos: los operadores pueden registrar períodos de alta transaccionalidad seguidos de lapsos de inactividad comercial, o exportar productos de naturaleza estacional. Esto supone un desafío para el modelamiento, ya que la construcción de una serie histórica para dicha combinación no siempre es posible. Frente a esta problemática, se implementó un esquema metodológico multidimensional que incorpora varias técnicas de *feature engineering*.

3.3.1. Ventana de observación e imputación jerárquica

Como se ha señalado, los datos se caracterizan por una alta intermitencia temporal; por ello, se definió una *ventana de observación* de 12 meses para cada combinación operativa. Bajo este criterio, si los antecedentes de una transacción registrada en enero de 2026 se ubican fuera del límite de esa ventana —esto es, antes de enero de 2025—, dicha transacción se clasifica como una *partida sin historia*. El uso de ventanas temporales dinámicas en el *feature engineering* se justifica por la necesidad de capturar la naturaleza cambiante de los datos, dado que la *normalidad* no es estática, sino que evoluciona con el tiempo. Una ventana de 12 meses permite capturar un ciclo estacional completo y garantiza la comparabilidad de las observaciones. En contraste, períodos más cortos podrían no representar adecuadamente la distribución de los datos, mientras que períodos más extensos podrían incorporar *deriva de concepto* al modelo (Park & Zamar, 2004).

No aplicar el criterio de ventana de observación constituye un error económico, ya que diversos factores exógenos, como choques de oferta, variaciones en los costos o cambios en la dinámica de los precios internacionales, pueden modificar la estructura de precios del producto. Por tanto, ignorar esta volatilidad del mercado internacional puede llevar a que el modelo incorpore datos obsoletos que no representan las condiciones vigentes del mercado de exportación y, en consecuencia, se alimente de información poco fiable.

Complementariamente, la mitigación de este problema se aborda bajo una lógica de *imputación jerárquica* aplicada a las variables construidas en esta fase. Este es un mecanismo de contingencia (*fall-back*) que funciona cuando la relación específica *exportador-país-partida* no satisface el criterio de *partida con historia*, el modelo escala automáticamente hacia un nivel superior de agregación, componiendo una relación alterna *exportador-partida*.

El objetivo de esta arquitectura de datos es preservar la representatividad del conjunto de entrenamiento (*training set*) y maximizar la densidad estadística de la información

suministrada al algoritmo. Bajo este diseño de validación multinivel, se dota al modelo de una mayor capacidad de generalización. Las variables creadas bajo esta lógica se definen a continuación:

- **Mediana del valor unitario:** La elección de la *mediana* como medida de tendencia central responde a su robustez frente a valores atípicos y ruido estadístico (Rousseeuw & Croux, 1993). En esta metodología se calcula mediante la lógica del *fallback* definida anteriormente.

$$\tilde{V}U_{e,p,k} = \text{mediana}(\{VU_t : (e, p, k) \text{ fijo}\}), \quad \tilde{V}U_{e,k} = \text{mediana}(\{VU_t : (e, k) \text{ fijo}\}) \quad (20)$$

La imputación jerárquica del valor de referencia se define como:

$$\widehat{V}U_{e,p,k} = \begin{cases} \tilde{V}U_{e,p,k}, & \text{si existe historia suficiente para } (e, p, k), \\ \tilde{V}U_{e,k}, & \text{en caso contrario.} \end{cases} \quad (21)$$

- **Z score:** Con el objetivo de cuantificar la dispersión relativa de los datos, se implementa una variante del *Z-score* robusta ante valores atípicos, basada en la Mediana y la Desviación Absoluta de la Mediana (*MAD*). De forma referencial, el *z-score* robusto original puede expresarse como:

$$Z_{VU} = \frac{VU - \tilde{M}}{1,4826 \cdot \text{MAD}(VU)} \quad (22)$$

Para el modelo propuesto, se aplican transformaciones logarítmicas sobre las diferencias absolutas con la finalidad de estabilizar la varianza en presencia de valores extremos y suavizar la escala de los residuos.

$$Z_{\log VU} = \frac{\ln(|VU - \tilde{M}| + 1)}{1,4826 \cdot \text{Mediana}(\ln(|VU - \tilde{M}| + 1))} \quad (23)$$

Donde⁴ \tilde{M} representa la *mediana de referencia*, la cual es imputada dinámicamente según la lógica de agregación jerárquica (base, país o partida) definida previamente.

Se aplica un factor de escala de 1,4826, propuesto por Rousseeuw y Croux (1993), para garantizar la consistencia estadística con una distribución normal y, de esta manera, hacer interpretables las desviaciones obtenidas.

⁴La diferencia se expresa en valor absoluto porque el logaritmo de un número negativo no está definido en los números reales. Asimismo, se suma 1 para evitar indeterminaciones cuando la diferencia es igual a cero.

Si bien, el *Z-score* es un indicador robusto de la dispersión de los datos, este puede representar fenómenos de inestabilidad numérica cuando las diferencias entre el valor unitario y la mediana son marginales. En tales casos, el denominador puede amplificar los valores artificialmente alcanzando magnitudes poco interpretables, que no necesariamente reflejan el verdadero alejamiento de las observaciones respecto de su trayectoria histórica.

Con el fin de mitigar este sesgo, y evitar introducir ruido al modelo, se realizó una técnica de *capping* o *acotamiento simétrico* de la variable, estableciendo un rango entre -15 y 15 . Este procedimiento asegura que los valores extremos de la variable no alteren la capacidad predictora del modelo.

$$Z_{\text{cap}} = \text{mín} (15, \text{máx} (-15, Z_{\log.VU})) \quad (24)$$

- **Delta del logaritmo del valor unitario:** Una de las ventajas del algoritmo *isolation forest* y de un diseño robusto de *feature engineering* es la capacidad de incorporar una *perspectiva histórica* en los datos. Esta variable compara el valor unitario actual con el valor unitario inmediatamente anterior identificado por el modelo dentro de la ventana de observación definida previamente, como se describe en la siguiente ecuación:

$$\Delta \ln(VU)_t = \ln(VU_t) - \ln(VU_{t-1}) \quad (25)$$

La transformación logarítmica del valor unitario permite capturar las variaciones relativas entre períodos. En este caso, las diferencias absolutas se convierten en una medida proporcional simétrica, que refleja qué tan grande es el valor actual frente al anterior, evaluando la magnitud del cambio en lugar de la diferencia nominal monetaria, la cual depende del tipo de producto.

3.3.2. Memoria histórica

Debido a que diversos productos de exportación están sujetos a una naturaleza estacional, es decir que responden a un comportamiento recurrente en meses específicos del año, es relevante implementar estrategias que sean robustas ante estas fluctuaciones cíclicas. Bajo este entendimiento se calcula la *mediana histórica (mensual o total)* y un *delta valor unitario*.

- **Mediana histórica:** Esta variable representa el estimador de tendencia central calculado únicamente sobre las observaciones correspondientes al mismo mes de años anteriores. Su objetivo principal es capturar la estacionalidad de la serie, permitiendo

que el modelo se ajuste y pueda diferenciar entre una anomalía real y un cambio cíclico esperado en los valores de la serie. Formalmente, para el mes m del año t , se define como:

$$\widetilde{VU}_{m,t} = \text{Mediana} \left(\{VU_{m,\tau}\}_{\tau \in \mathcal{T}_m, \tau < t} \right) \quad (26)$$

donde $VU_{m,\tau}$ es el valor unitario observado en el mes m del año τ , y \mathcal{T}_m representa el conjunto de años disponibles en la base para ese mes.

En este caso, también se aplica una lógica de imputación jerárquica y de contingencia. Si el modelo no encuentra valores para el mes m en ningún año disponible, se calcula la *mediana histórica total*, definida como la mediana de todas las observaciones registradas para la relación *exportador-país-partida*:

$$\widetilde{VU}^{\text{total}} = \text{Mediana} \left(\{VU_i\}_{i \in \mathcal{I}} \right) \quad (27)$$

donde \mathcal{I} es el conjunto de observaciones históricas disponibles para dicha relación.

- **Delta del valor unitario:** Esta variable se define como la diferencia entre el valor unitario observado en el mes actual y la mediana histórica obtenida mediante el proceso de *fallback*. Si la *mediana histórica mensual* está disponible, la comparación se realiza respecto de esta; en caso contrario, se utiliza la *mediana histórica total* para efectuar el cálculo:

$$\Delta VU_{m,t} = VU_{m,t} - \widetilde{VU}_{m,t}^*, \quad \widetilde{VU}_{m,t}^* = \begin{cases} \widetilde{VU}_{m,t}, & \text{si existe mediana histórica mensual,} \\ \widetilde{VU}^{\text{total}}, & \text{en caso contrario.} \end{cases} \quad (28)$$

3.4. Datos

El algoritmo fue entrenado con un gran volumen de información, lo cual es crucial para que el modelo aprenda patrones de comportamiento, relaciones y asociaciones en los datos. Como señalan Sivakumar et al. (2024), mientras más información contenga el *training set*, mayor será la capacidad del modelo para generalizar y aplicar ese conocimiento sobre el *testing set*. La distribución del valor unitario utilizada en este análisis se presenta en el anexo F.

Como se observa en el anexo G, la transformación logarítmica del valor unitario converge hacia una distribución aproximadamente gaussiana, lo que indica que los registros

anómalos se ubican en los márgenes de la distribución. Esto significa que las anomalías se aíslan de las transacciones comunes en las colas de la distribución.

3.5. Implementación de Isolation Forest

La fase de implementación del *Isolation Forest* operacionaliza la arquitectura conceptual definida en el apartado teórico de la investigación y en las fases metodológicas previas, hacia los registros administrativos de exportaciones con la finalidad de asignar un *Score de Anomalía* a cada transacción del set de enero 2026, utilizando la estructura de los "bosques" para generar particiones basadas en los *Features* definidos con anterioridad.

Para implementar el *Isolation Forest* se utilizó el software especializado *RStudio*, debido a sus ventajas para el procesamiento del modelo. En particular, se empleó el paquete *isotree* desarrollado por Cortes (2026)⁵, que implementa la formulación base del *Isolation Forest* (Liu et al., 2008) y su extensión *Extended Isolation Forest* (Hariri et al., 2021).

Dentro del paquete *isotree*, la librería usada fue *isoforest*, la cual posibilita la aplicación del *Isolation Forest* sobre los datos. Para este ejercicio se implementaron 4 variantes al algoritmo principal de *Isolation Forest*, el *Extended Isolation Forest*, *Faircut Forest* y *Robust Random-cut Forest*⁶.

Tras la revisión de la literatura especializada y la naturaleza de los registros administrativos, se seleccionó el *Extended Isolation Forest (EIF)* como algoritmo principal para la detección de anomalías en los registros de comercio exterior. Esta elección se fundamenta en su mayor robustez frente a la elevada heterogeneidad del valor unitario y en su capacidad para mitigar los efectos de *swamping* y *masking*, fenómenos que pueden sesgar la clasificación del algoritmo. Esta decisión se corroboró mediante métricas de desempeño, que mostraron que el EIF fue la técnica más robusta para detectar anomalías en las operaciones de exportación.

Antes de explicar la parametrización del modelo principal y las otras variantes del mismo, es necesario definir las variables de las cuales se alimentan los algoritmos.

3.5.1. Definición de los atributos

La selección adecuada de los atributos del modelo es crucial, ya que de ella depende directamente su capacidad de clasificación. Como señala Ski (2024), la selección de atributos tiene un impacto significativo en el poder predictivo del modelo; por ello, estos deben definirse cuidadosamente.

⁵Disponible en el repositorio oficial de CRAN: <https://cran.r-project.org/web/packages/isotree/index.html>

⁶la implementación de las distintas variantes del modelo corresponden a cambios en los parámetros del algoritmo base

En la presente investigación, el vector de atributos se conformó a partir de las variables que fueron creadas mediante el proceso de la ingeniería de atributos (*feature engineering*), las que se detallan en la siguiente tabla:

Cuadro 1: **Definición del vector de atributos (Features) para el EIF**

Atributo	Descripción	Justificación técnica
Me_{sin}	Seno del componente mensual	Codifica cíclicamente el tiempo para preservar la proximidad entre diciembre y enero, y captura la estacionalidad anual.
Me_{cos}	Coseno del componente mensual	Complementa la codificación cíclica temporal para representar la estacionalidad de forma continua.
Partida sin historia	Indicador binario de ausencia de historia.	Identifica registros sin historia y previene la introducción de sesgos en el cálculo del puntaje.
Z_{cap}	Puntaje <i>Z-score</i> del registro.	Mide la desviación del valor unitario respecto de la mediana; aplica una transformación logarítmica y un <i>capping</i> en $[-15; 15]$ para neutralizar el ruido.
$\Delta \ln(VU)_t$	Diferencia logarítmica entre el <i>VU</i> del registro y su inmediato anterior.	Añade un componente de control histórico y tendencial al modelo al incorporar el cambio mensual.
$\Delta VU_{m,t}$	Diferencia entre el <i>VU</i> del registro y la mediana resultante del <i>fallback</i> . ⁷	Mide la desviación absoluta del <i>VU</i> frente al valor de referencia de mercado; añade una perspectiva histórica y captura anomalías en su escala original, complementando el enfoque logarítmico.

Fuente: Elaboración propia.

Los atributos no fueron elegidos al azar; se evaluó el desempeño de distintas combinaciones de variables y transformaciones antes de definir el vector de atributos final. Asimismo, esta selección se fundamenta metodológicamente en la necesidad de capturar, de forma multidimensional, las anomalías en el comercio exterior. Por ello, se integran variables que reflejan la magnitud y proporción de los cambios, la consistencia histórica mediante lógicas de *fallback*, así como patrones estacionales y tendenciales. De esta forma,

el algoritmo dispone de una amplia base de información para optimizar la detección de registros anómalos.

3.5.2. Parametrización de los algoritmos

Así como es relevante definir correctamente el vector de atributos, también lo es establecer adecuadamente los parámetros con los que se construye el modelo, ya que de ellos depende la variante del IF utilizada y, en consecuencia, los resultados obtenidos.

Cuadro 2: Configuración técnica de los modelos de aislamiento evaluados

Parámetro	IF Estándar	Extended IF	Fair Cut IF	Robust IF
<i>ntrees</i>	200	200	200	200
<i>sample_size</i>	256	256	256	256
<i>ndim</i> (Cortes)	1	\sqrt{p}	1	1
<i>coefs</i>	-	Normal	Normal	-
<i>ntry</i>	-	\sqrt{p}	1	-
<i>prob_avg_gain</i>	0	0	0	1
Enfoque	Particiones eje-paralelas	Hiperplanos aleatorios	Balanceo de cortes	Ganancia de información

Fuente: Elaboración propia basada en la implementación del paquete *isotree*.

Respecto de la parametrización del modelo *Extended Isolation Forest*, la configuración responde a una calibración precisa de su arquitectura interna, orientada a la estabilidad estadística. En primer lugar, el número de árboles se fijó en $ntrees = 200$ y el tamaño de muestra en $sample_size = 256$, con el fin de garantizar la convergencia del puntaje de anomalía y mitigar los efectos de *swamping* y *masking*. Por otro lado, la dimensionalidad de las particiones se estableció en $ndim = \sqrt{p}$ ⁸. Esta decisión metodológica se fundamenta en la necesidad de reforzar la robustez del modelo frente a la maldición de la dimensionalidad (Liu et al., 2012). Con ello, se incrementa la diversidad estructural del bosque al forzar a los árboles individuales a utilizar combinaciones aleatorias de atributos. De esta manera, se evita que atributos ruidosos o con magnitudes desproporcionadas sesguen la capacidad de detección de anomalías del algoritmo, y se favorece que las anomalías se aislen por las propiedades del modelo y no por sesgos geométricos del espacio global. Para mejorar su precisión, se optimizó la selección de cortes mediante $ntry = \sqrt{p}$; este parámetro evalúa múltiples candidatos de corte en cada nodo del bosque (generación de hiperplanos) y selecciona aquel que maximiza la ganancia de información local.

⁸Si bien Hariri et al. (2021) sugieren utilizar el total de dimensiones para generar una estructura global, en esta investigación se emplea un submuestreo de atributos como estrategia de regularización frente al posible *sobreajuste* del modelo

Una vez calibrados los parámetros del *Extended Isolation Forest*, los cuales permiten optimizar la detección de anomalías en los registros de comercio exterior, resulta pertinente explicar cómo se estableció el *threshold*, o *umbral*, que permite clasificar un registro como anómalo o normal. A continuación, se detalla el procedimiento seguido y las justificaciones metodológicas utilizadas para definir este límite.

3.5.3. Especificación del umbral de clasificación

Como señalan Tamakloe et al. (2025), definir correctamente un umbral de anomalía constituye un aspecto crítico del modelo. Dado que de ello dependen los resultados que este arroja, un umbral mal seleccionado puede generar un exceso de falsos positivos o permitir que anomalías reales sean clasificadas como datos normales. En este contexto, debe establecerse un umbral que permita clasificar los registros según su puntaje de anomalía; los autores recomiendan que un experto determine dicho punto de corte. Sin embargo, al tratarse de un método manual, este enfoque resulta subjetivo y poco práctico para un sistema automatizado.

Originalmente, Liu et al. (2008) definían su umbral de anomalía en 0,60⁹. Sin embargo, la definición del umbral clasificador depende de las características de los datos y la naturaleza del *score* predicho, por lo que para este ejercicio se definió el umbral en 0,70. La regla de clasificación se define de la siguiente forma:

$$\hat{y}_i = \begin{cases} 0, & \text{si } s_i < 0,70 \\ 1, & \text{si } s_i \geq 0,70 \end{cases} \quad (29)$$

donde s_i es el *score* de anomalía del registro i , $\hat{y}_i = 0$ indica un registro normal y $\hat{y}_i = 1$ indica un registro anómalo.

La determinación del umbral crítico para este ejercicio se realizó mediante un proceso de tres etapas. En primer lugar, se llevó a cabo un análisis de sensibilidad¹⁰ para evaluar la estabilidad del modelo. Este ejercicio consistió en examinar iterativamente distintos umbrales dentro del rango [0,30; 0,90], con incrementos de 0,01. Ello permitió observar el número de anomalías capturadas para cada valor del umbral, así como identificar un punto de corte que estabilizara la capacidad predictiva del modelo. En este contexto, el umbral óptimo es aquel que minimiza los falsos positivos y asegura que la clasificación responda a anomalías efectivas, y no a la volatilidad de los datos o de los atributos.

Como se observa en el anexo H, el número de registros clasificados como anómalos se

⁹Cabe mencionar que los autores de la investigación mencionada definieron este umbral para un ejercicio realizado con una base de datos sintética

¹⁰Las pruebas de sensibilidad del umbral se realizaron sobre las predicciones del *Extended Isolation Forest* definido anteriormente.

reduce de manera marcada entre los umbrales 0,35 y 0,60, y tiende a estabilizarse a partir de 0,70. Este comportamiento respalda la selección de dicho valor como punto de corte, en la medida en que permite reducir la presencia de falsos positivos sin comprometer la detección de anomalías relevantes.

En segundo lugar, este umbral se validó mediante juicio experto, al contrastar el conocimiento sobre la dinámica de los registros aduaneros con el análisis de los puntajes predichos por el modelo.

Finalmente, para dotar al procedimiento de mayor robustez y objetividad, se aplicaron diversas pruebas de validación estadística que respaldaron la pertinencia del valor seleccionado. Para evaluar la estabilidad distribucional del umbral, se analizó su posición con respecto al primer y segundo momento de la distribución de los puntajes de anomalía. En primer lugar, se calcularon la media aritmética (μ) y la desviación estándar (σ) de las observaciones. A partir de estos parámetros, se realizó una simulación de Monte Carlo basada en 10.000 iteraciones de una distribución nula con las características de los datos observados, $N(\mu, \sigma^2)$. Los resultados de este ejercicio muestran que la probabilidad de que un puntaje supere el umbral crítico definido debido únicamente a fluctuaciones aleatorias es inferior al 0,1 %. Por lo tanto, el umbral seleccionado se ubica en una de las colas extremas de la distribución, lo que respalda que las anomalías detectadas correspondan a observaciones estructuralmente atípicas y relevantes para el monitoreo continuo de las operaciones de exportación.

4. RESULTADOS

Una vez ejecutado el algoritmo de *Isolation Forest* y sus variantes, los resultados obtenidos de la clasificación se resumen en la siguiente tabla, construida sobre un total de 15.055 observaciones.

Como se observa en la Tabla 3, la proporción de registros anómalos es baja, correspondiendo a menos del 1 % del total de la muestra. Por lo tanto, se cumple con la definición de anomalía propuesta por Liu et al. (2012), quienes establecen que estas observaciones son escasas y distintas, por lo que pueden aislarse con mayor facilidad.

Cuadro 3: **Comparación de anomalías detectadas por variante del modelo**

Modelo	Anomalías	Porcentaje de anomalías
EIF	94	0,62 %
Robust	67	0,45 %
Fair Cut	50	0,33 %
Estándar	16	0,11 %

Fuente: Elaboración propia.

Adicionalmente, estos resultados son consistentes con la teoría de los bosques de aislamiento respecto del desempeño de sus distintas variantes. En particular, el *Extended Isolation Forest* es el modelo que identifica el mayor número de anomalías, seguido por el *Robust Isolation Forest*, el *Fair Cut Isolation Forest* y, finalmente, la especificación estándar. Este patrón sugiere que la variante *EIF* ofrece una mayor capacidad de detección en este conjunto de datos, al identificar observaciones atípicas que otras especificaciones no logran aislar.

Dado el enfoque no supervisado del algoritmo, se definió un *Ground Truth* mediante un proceso de validación de los resultados del modelo realizado por expertos. El proceso de validación de resultados se resume en el anexo M.

Para evaluar y comparar el desempeño de los distintos modelos, es necesario contar con métricas de validación. Tamakloe et al. (2025) señalan que, en modelos de clasificación, suele recurrirse a indicadores como el *recall*, la *precisión* y el *F1-score*. Sin embargo, dado que en este caso se trata de un modelo no supervisado y de una base de datos de gran volumen, una de las principales limitaciones fue el cálculo conjunto de estas métricas para el total de *test set*. Por ello, en este ejercicio únicamente se estimó la precisión para todos los modelos, la cual se calculó de la siguiente manera:

$$\text{Precisión} = \frac{\text{Tasa de verdaderos positivos}}{\text{Tasa de verdaderos positivos} + \text{Tasa de falsos positivos}} \quad (30)$$

Los resultados se presentan en la siguiente tabla:

Cuadro 4: Precisión por variante del modelo

Modelo	Anomalías	Falsos positivos	Verdaderos positivos	Precisión
EIF	94	17	77	81,91 %
Robust	67	26	41	61,19 %
Faircut	50	11	39	78,00 %
Estandar	16	0	16	100,00 %

Fuente: Elaboración propia.

Estos hallazgos respaldan el uso del *EIF* como el algoritmo principal para la generación de alertas en los registros de exportación de comercio exterior. Si bien la especificación estándar alcanza una precisión de 100 %, identifica únicamente 16 anomalías, lo que sugiere que su capacidad de detección de anomalías es muy limitada. En contraste, el *Extended Isolation Forest* no solo detecta el mayor número de anomalías, sino que también tiene una precisión alta; esto indica que efectivamente la configuración del modelo y sus atributos permiten que se detecten las anomalías relevantes y se ignore el ruido estadístico.

Por su parte, el *Fair Cut Isolation Forest* presenta un alto desempeño, menor que el *Extended Isolation Forest* y superior a la del *Robust Isolation Forest*; sin embargo, detecta un menor número de anomalías que el primero. Si bien el *Robust Isolation Forest* es el

segundo modelo que captura una mayor proporción de anomalías, su precisión sugiere que estas corresponden a falsos positivos. En conjunto, los resultados presentados demuestran que el modelo que mejor balance entre precisión y capacidad de detección de anomalías es el *Extended Isolation Forest*.

Dado que el *ground truth* definido no cuenta con un etiquetado previo que permita realizar una validación cruzada convencional (k-fold, leave-one-out) no es posible calcular el *recall* sobre el total de la base de datos, por lo tanto, se realizó un experimento de validación externa. Este consistió en inyectar un conjunto de anomalías controladas sobre una muestra y evaluar la capacidad de detección del modelo. En particular, se seleccionó una muestra representativa de 375 observaciones previamente definidas y validadas como normales; posteriormente, 50 de estas observaciones fueron perturbadas con anomalías en el valor unitario y el modelo se ejecutó sobre esta base de datos sintética.

Como se observa en el anexo N, de las 50 anomalías inducidas el modelo identificó correctamente 30, lo que representa un *recall* o *sensibilidad* del 60 %. Cabe destacar que el modelo no generó falsos positivos dentro del grupo de 325 registros normales; por tanto, su especificidad fue del 100 %. Esto confirma, además, que el umbral de anomalía fue adecuadamente definido y que el modelo es robusto frente a las anomalías detectadas. En conjunto, estos resultados muestran que, aunque el modelo es conservador y tiende a ignorar anomalías marginales, posee una alta eficacia para detectar desviaciones estructurales que representan variaciones atípicas significativas.

En el anexo J se observa la relación entre el logaritmo del valor unitario y el puntaje de anomalía asignado por el modelo. La mayor parte de las operaciones se concentra en niveles de anomalía inferiores a 0,50, lo que corresponde, en general, a registros comerciales normales. En contraste, las observaciones clasificadas como anómalas se ubican por encima del umbral crítico de 0,70. Cabe destacar que las anomalías detectadas no se limitan a valores unitarios extremos, sino que corresponden a observaciones que rompen con su comportamiento histórico o con la dinámica esperada del mercado.

Como se observa en el anexo I, la distribución de los puntajes del modelo evidencia una clara concentración de registros en el intervalo [0,35; 0,60]. Esta configuración permite distinguir las anomalías estructurales, ubicadas en el extremo derecho de la distribución, las cuales son aisladas a partir del umbral crítico definido y validado previamente.

4.1. Pruebas post-estimación y explicabilidad del modelo

Dado que uno de los principales inconvenientes de los modelos de aprendizaje automático es la explicación de su lógica interna y su interpretabilidad, para este ejercicio se aplicaron dos pruebas post-estimación: *Local Interpretable Model-Agnostic Explanations* (LIME) (Ribeiro et al., 2016) y *Permutation Feature Importance* (Altmann et al., 2010; Breiman, 2001). La primera metodología permite evaluar localmente la decisión del mo-

delo, mientras que la segunda sirve para identificar cuales son las variables más relevantes mediante la permutación de los atributos.

Según los resultados de LIME (revisar anexo K) las variables que más inciden en la detección de anomalías son $\Delta VU_{m,t}$, $\Delta \ln(VU)_t$ y, en menor medida, Z_{cap} . Por otro lado, *Partida sin historia*, Mes_{sin} y Mes_{cos} no presentan grandes aportes, por lo cual el modelo se basa en las desviaciones estructurales del valor estacionario más que en componentes estacionales.

Los resultados del ejercicio de permutación de atributos que se observa en el anexo L, confirman lo explicado por LIME. Como se evidencia, la mayor caída en la capacidad de detección de anomalías se observa cuando se permutan $\Delta VU_{m,t}$ y $\Delta \ln(VU)_t$, mientras que Z_{cap} también presenta un efecto relevante pero menor¹¹ Sin embargo, cuando se permutan Mes_{sin} y Mes_{cos} no existe cambio en el desempeño del modelo, y la de *Partida sin historia* incrementa el conteo de anomalías, esto sugiere que estas variables cumplen un rol contextual más que productivo para el modelo.¹² De forma conjunta, ambas pruebas confirman que la capacidad de detección del modelo está determinada por variables que capturan principalmente las alteraciones del valor unitario.

Respecto del riesgo de sobreajuste, el EIF incorpora regularización estructural mediante submuestreo aleatorio ($\psi = 256 \ll n = 703.169$), lo que impide que cualquier árbol individual memorice el conjunto de entrenamiento (Liu et al., 2012). La estabilidad del *score* ante distintas semillas aleatorias fue verificada mediante 20 réplicas independientes del modelo, con una tasa de concordancia en la clasificación binaria superior al 95 %.

5. LIMITACIONES

Las principales limitaciones del trabajo se relacionan con las características de la información disponible. Los registros de las transacciones comerciales suelen ser intermitentes; sin embargo, esta condición no responde necesariamente a una baja calidad de la fuente ni a deficiencias en el proceso de recopilación de datos, sino a la propia naturaleza del comercio internacional. En efecto, un operador no necesariamente exporta el mismo producto todos los meses, y los flujos comerciales pueden estar sujetos a fluctuaciones estacionales o a *shocks* de demanda que explican dicha intermitencia. No obstante, esta característica dificulta la construcción de series históricas completas para la combinación *exportador-país-partida*, lo que complejiza la definición de qué debe considerarse una anomalía específica para cada combinación.

Otra limitación se relaciona con la granularidad de los datos. En los registros ad-

¹¹Estas tres variables resultaron estadísticamente significativas al 1 % al aplicar una permutación Monte Carlo con 100 iteraciones. El p-valor se calculó como la probabilidad de obtener, tras la permutación, un número de anomalías igual o superior al observado sin permutar.

¹²Estas tres variables no resultaron significativas a los niveles de confianza analizados.

ministrativos de la SENAE, el nivel más desagregado de información corresponde a la Declaración Aduanera de Exportación (DAE), que asigna un código único a cada *ítem* dentro de una misma operación comercial. Dado que este identificador tiene un carácter estrictamente transaccional, cambia en cada nueva operación, lo que impide construir una serie histórica continua para un mismo flujo, incluso cuando un mismo exportador envía exactamente el mismo producto en distintos momentos. Por esta razón, la metodología desarrollada no pudo aplicarse al nivel más bajo de desagregación, sino a un nivel de agregación superior.

6. CONCLUSIONES Y RECOMENDACIONES

En conclusión, los resultados obtenidos a partir de la implementación del algoritmo *Extended Isolation Forest* permiten afirmar que el uso de técnicas de aprendizaje automático representa una mejora sustancial en la fiscalización de registros administrativos. La robustez del modelo se valida mediante las pruebas post-estimación, su significancia estadística y su desempeño general, los cuales muestran una jerarquía clara en la definición de una *anomalía*. En particular, $\Delta VU_{m,t}$ y $\Delta \ln(VU)_t$ constituyen los principales pilares de la capacidad de aislamiento del algoritmo, con reducciones aproximadas del 77,6 % y 56,38 % en la detección de anomalías cuando estas variables son permutadas. Este hallazgo confirma que el valor unitario actúa como una señal central para identificar inconsistencias en los registros administrativos y sugiere que las anomalías en el comercio exterior ecuatoriano se manifiestan como quiebres estructurales en los valores declarados, es decir, como desviaciones atípicas respecto del comportamiento esperado.

Asimismo, orientar este ejercicio hacia un marco de inteligencia artificial explicable (XAI), mediante la integración de técnicas como LIME y la importancia por permutación, refuerza la validez de los resultados obtenidos, ya que garantiza que el algoritmo no funcione como una «caja negra», sino como una herramienta transparente y explicable para comprender las decisiones del modelo en el contexto de los registros de comercio exterior. Este enfoque reduce la incertidumbre en la formulación de los puntajes de riesgo y facilita la toma de decisiones informadas en los procesos de auditoría y control aduanero. En suma, el modelo propuesto optimiza la identificación de registros anómalos bajo un umbral crítico superior a 0,70 y, además, ofrece un marco metodológico innovador, robusto, escalable y auditable que contribuye a fortalecer la integridad de las estadísticas de comercio exterior.

A partir de los resultados presentados, se recomienda que futuras investigaciones realicen un *benchmark* para comparar el desempeño y la eficacia del *Extended Isolation Forest* frente a otros algoritmos de *machine learning*. En particular, sería pertinente contrastarlo con modelos basados en densidad, como *Local Outlier Factor* y *DBSCAN*; con modelos basados en distancia, como *k-Nearest Neighbors* y *ORCA*; con enfoques de clasificación,

como *One-Class SVM*; y con métodos de aprendizaje profundo, como *Autoencoders* y *Variational Autoencoders*. Asimismo, convendría profundizar en otras variantes de la familia *Isolation Forest* que no fueron consideradas en este estudio, con el fin de identificar posibles mejoras en la capacidad predictiva del modelo.

Por otro lado, sería fundamental incorporar algoritmos de Procesamiento de Lenguaje Natural (PLN) para aumentar la precisión de las auditorías sobre los registros de comercio exterior. Estas herramientas permitirían agrupar variedades de producto a partir de la descripción comercial y técnica contenida en las declaraciones aduaneras, con el objetivo de alcanzar un nivel de granularidad superior al utilizado en este estudio y, de ese modo, reducir el ruido generado por la heterogeneidad de los productos dentro de una misma partida. Adicionalmente, para reforzar la transparencia del modelo, se recomienda incorporar SHAP (*SHapley Additive exPlanations*) como una técnica complementaria que fortalezca la explicabilidad global del algoritmo y profundice el enfoque de Inteligencia Artificial Explicable.

REFERENCIAS

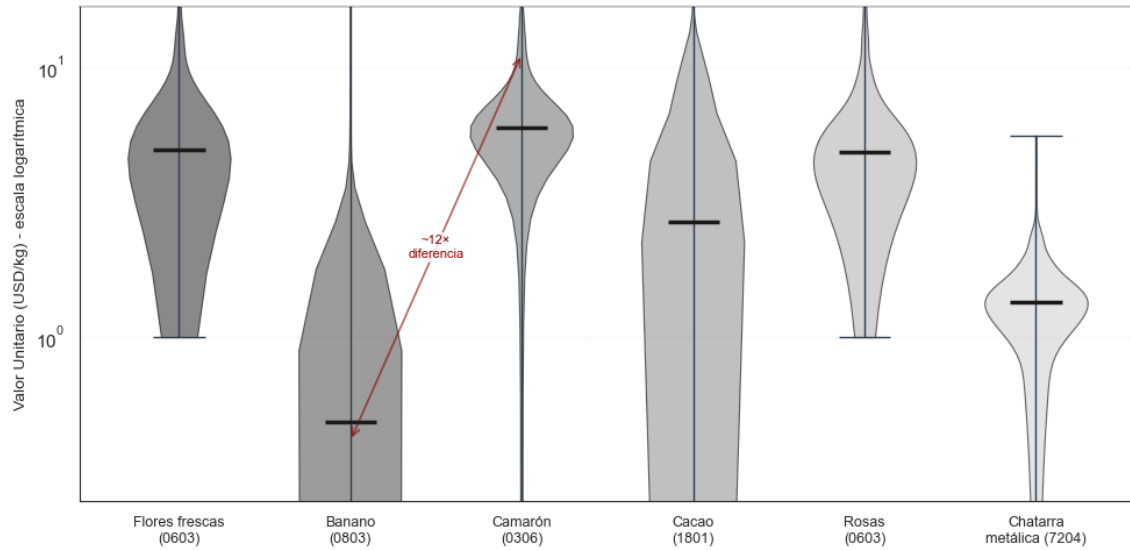
- Aggarwal, C. C. (2017). *Outlier Analysis* (2.^a ed.). Springer. <https://doi.org/10.1007/978-3-319-47578-3>
- Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10), 1340-1347. <https://doi.org/10.1093/bioinformatics/btq134>
- Ayoub, M., Khalid, J., & Karczmarek, P. (2022). FUZZY C-MEANS based extended isolation forest for anomaly detection. *International Conference on Advanced Intelligent Systems for Sustainable Development*, 411-418. https://doi.org/10.1007/978-3-031-02447-4_43
- Baffes, J., & Ajwad, M. I. (2001). Identifying Price Linkages: A Review of the Literature and an Application to the World Market of Cotton. *Applied Economics*, 33(15), 1927-1941.
- Bandaragoda, T., Ting, K. M., Albrecht, D., Liu, F. T., & Wells, J. R. (2014). Efficient Anomaly Detection by Isolation Using Nearest Neighbor Ensemble. *2014 IEEE International Conference on Data Mining Workshop*, 698-705.
- Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3), 1-58. <https://doi.org/10.1145/1541880.1541882>
- Cortes, M. D. (2026). Package 'isotree'.
- Hariri, S., Kind, M. C., & Brunner, R. J. (2021). Extended Isolation Forest. *IEEE Transactions on Knowledge and Data Engineering*, 33(4), 1479-1489. <https://doi.org/10.1109/tkde.2019.2947676>
- Hinkle, D. E., & Herrou-Aramburu, A. (1994). *Unit Values as Proxies for Export Prices: An Empirical Analysis* (Working Paper). World Bank.
- Hubert, M., & Vandervieren, E. (2008). An Adjusted Boxplot for Skewed Distributions. *Computational Statistics & Data Analysis*, 52(12), 5186-5201. <https://doi.org/10.1016/j.csda.2007.11.008>
- International Monetary Fund, International Labour Organization, Organisation for Economic Co-operation and Development, Statistical Office of the European Communities (Eurostat), United Nations Economic Commission for Europe & The World Bank. (2009). *Export and Import Price Index Manual: Theory and Practice* [Includes bibliographical references and index].

- Kuhn, M., & Johnson, K. (2020). Feature Engineering and Selection: A Practical Approach for Predictive Models. *The American Statistician*, 74(3), 308-309. <https://doi.org/10.1080/00031305.2020.1743656>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413-422. <https://doi.org/10.1109/ICDM.2008.17>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data*, 6(1). <https://doi.org/10.1145/2133360.2133363>
- Miljković, D. (2010). Review of novelty detection methods, 593-598.
- Moomtaheen, F., Bagui, S. S., Bagui, S. C., & Mink, D. (2024). Extended isolation forest for intrusion detection in Zeek data. *Information*, 15(7), 404. <https://doi.org/10.3390/info15070404>
- Páramo, J. C. (2019). *Unit Values and Price Indices for Exports: Estimation and Uses in the Context of Ecuador's Central Bank* (Working Paper). Banco Central del Ecuador.
- Park, J.-P., & Zamar, R. (2004). The Detection and Testing of Multiple Outliers in Linear Regression. *Journal of the Korean Data and Information Science Society*, 15.
- Rajeev, H., & Devi, U. (2022). Detection of credit card fraud using isolation forest algorithm. En *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021* (pp. 23-34). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-5640-8_3
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR*, *abs/1602.04938*. <http://arxiv.org/abs/1602.04938>
- Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424), 1273-1283. <https://doi.org/10.1080/01621459.1993.10476408>
- Shao, M., Shao, H., Wang, X., Gao, Y., & Liu, B. (2025). Interpretable anomaly detection using extended isolation forest with adaptive thresholds [Publicado en línea antes de impresión]. *Structural Health Monitoring*. <https://doi.org/10.1177/14759217251339607>
- Silver, M. (2007). *Do Unit Value Export, Import, and Terms of Trade Indices Misrepresent Price Changes? A Problem of Heterogeneity* (IMF Working Paper N.º WP/07/122). International Monetary Fund.
- Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). Trade-off between training and testing ratio in machine learning for medical image processing. *PeerJ Computer Science*, 10, e2245.
- Ski, M. C. (2024). An overview of outlier detection methods. *London J. Eng. Res*, 24(2), 37-79.

- Tamakloe, K., Bonsu, G., Chaganti, S. K., Sheikh, A., & Chen, D. (2025). Automated Detection of Site-to-Site Variations: A Sample-Efficient Framework for Distributed Measurement Networks. *Eng*, 6(11). <https://doi.org/10.3390/eng6110297>
- United Nations Statistics Division (UNSD). (2019, junio). *Trade Data Processing and Validation Methodology* (inf. téc.) (Part: UN Comtrade. Detecta extremos comparando el valor unitario declarado con el estandar (SUV) usando limites $SUV \times 100$ y $SUV / 100$). United Nations, Department of Economic y Social Affairs. https://comtradeapi.un.org/files/v1/app/wiki/UNSD_Method_trade_data_processing_v6-17_Jun_2019.pdf
- Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. *Statistical Analysis and Data Mining*, 5(5), 363-387. <https://doi.org/10.1002/sam.11161>

ANEXOS

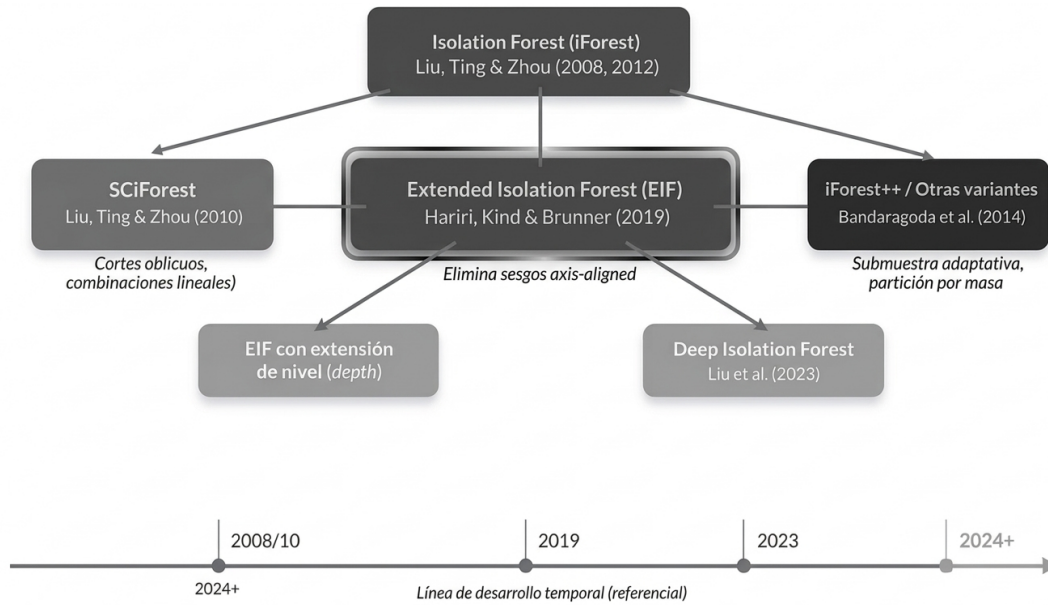
ANEXO. A HETEROGENEIDAD DEL VALOR UNITARIO ENTRE CATEGORÍAS DE EXPORTACIÓN



Fuente: Elaboración propia.

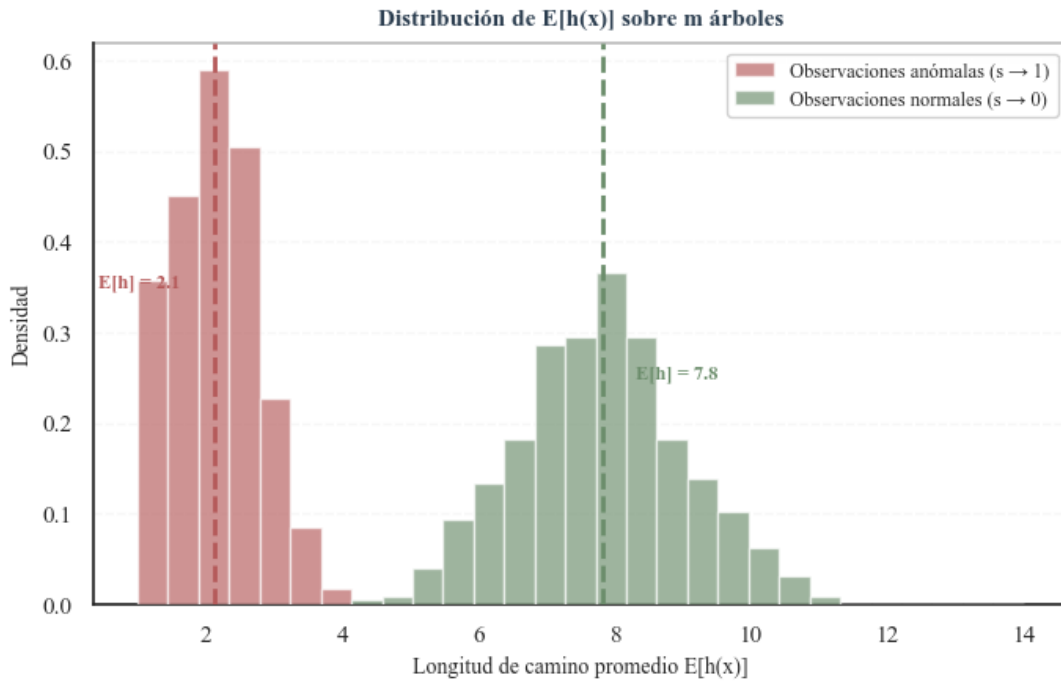
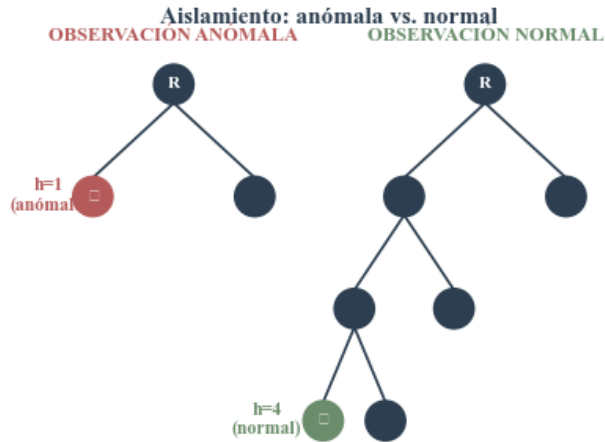
Nota. Heterogeneidad en las escalas del valor unitario entre categorías de exportación ecuatorianas. Un único umbral absoluto de rareza no puede funcionar correctamente cuando las distribuciones difieren en órdenes de magnitud entre categorías.

ANEXO. B FAMILIA DE ALGORITMOS ISOLATION FOREST



Fuente: Elaboración propia.

ANEXO. C PRINCIPIO DE AISLAMIENTO



Fuente: Elaboración propia.

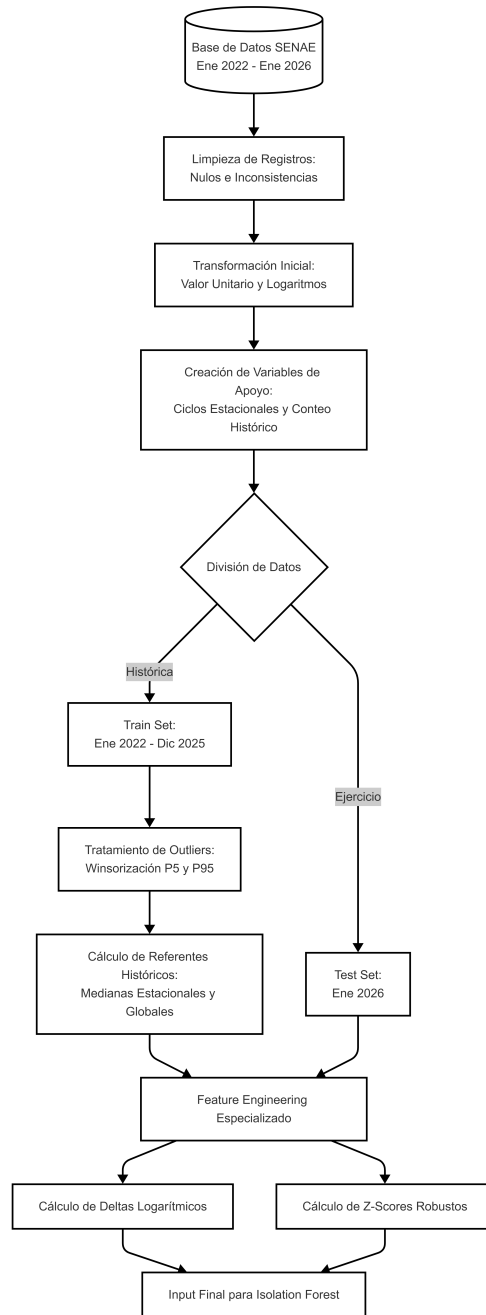
El principio de aislamiento. Arriba: un árbol aísla a una observación anómala en menos pasos, mientras que una normal requiere muchos más. Abajo: distribución de longitudes de camino promedio $E[h(x)]$ sobre m árboles; la separación entre normales y anómalas fundamenta el score $s(x)$.

ANEXO. D IMPACTO DE LA WINSORIZACIÓN

Categoría	Total	Porcentaje (%)
Total de datos	703.169	100 %
Datos sin cambios	578.001	82 %
Reemplazo P_5	62.393	9 %
Reemplazo P_{95}	62.775	9 %

Fuente: Elaboración propia.

ANEXO. E FLUJOGRAMA DEL TRATAMIENTO DE LAS BASES DE DATOS



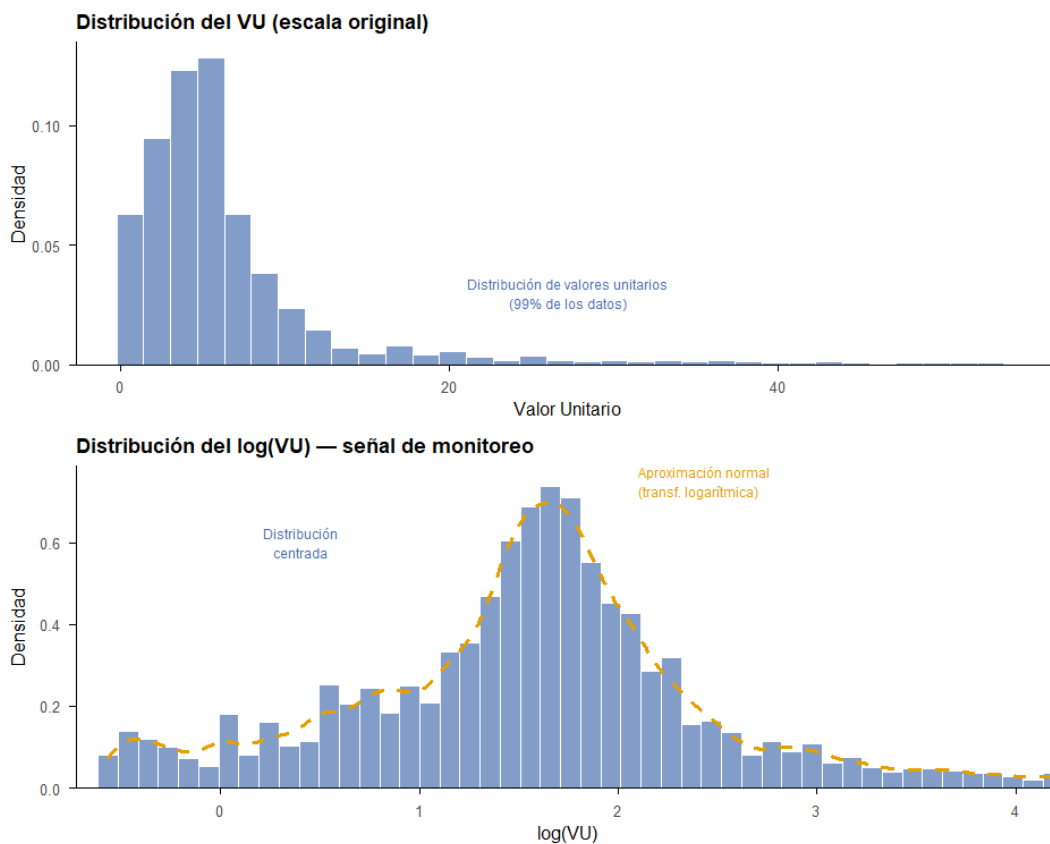
Fuente: Elaboración propia

ANEXO. F VOLUMEN DE OBSERVACIONES DE LAS BASES UTILIZADAS

Base	N. Observaciones	Periodo
Base histórica (<i>Train set</i>)	703.169	Enero 2022 – Diciembre 2025
Base de ejercicio (<i>Test set</i>)	15.055	Enero 2026

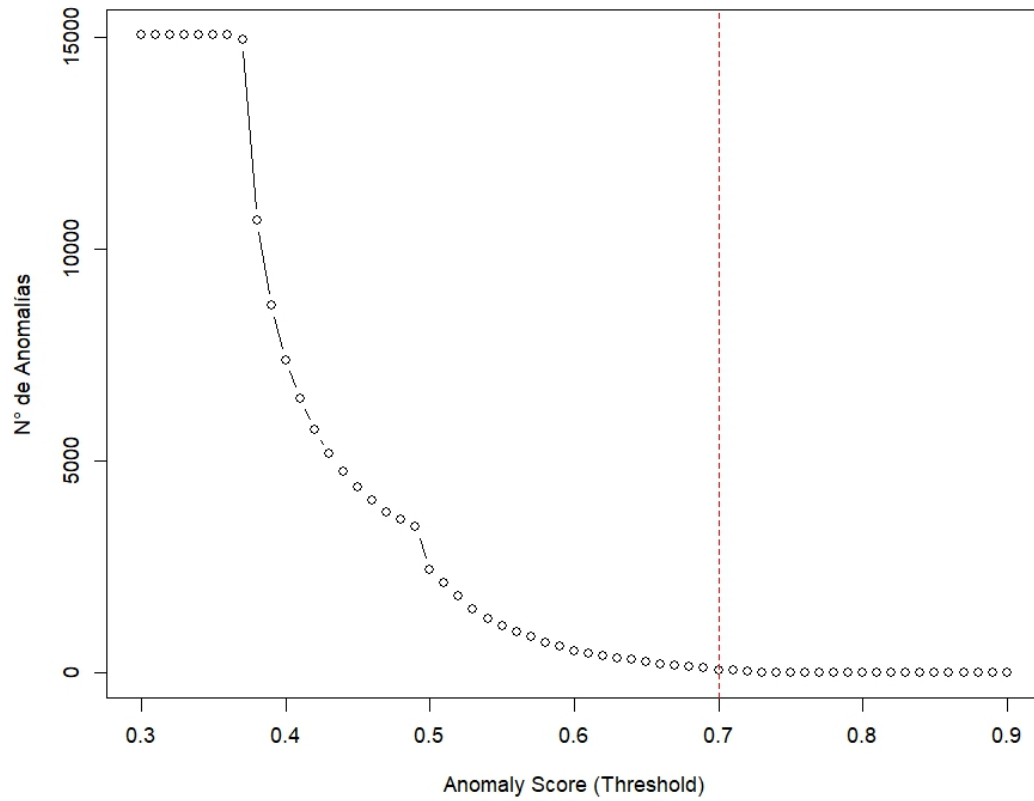
Fuente: Elaboración propia.

ANEXO. G EFECTO DE LA TRANSFORMACIÓN LOGARÍTMICA DEL VALOR UNITARIO



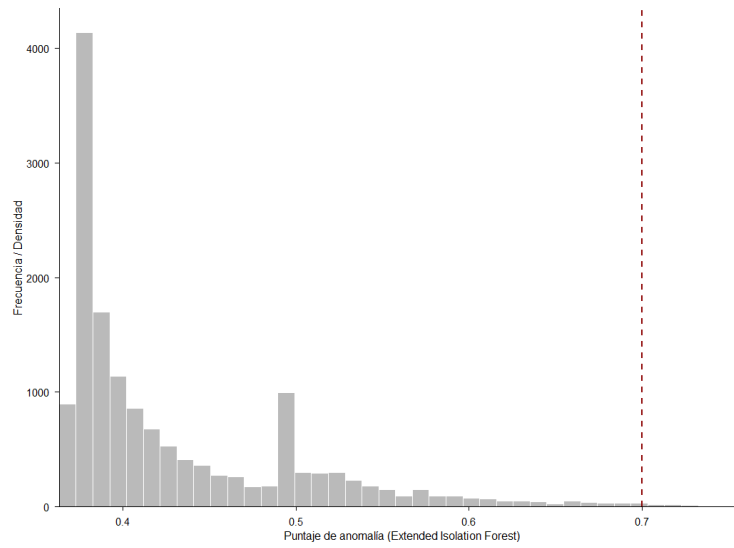
Fuente: Elaboración propia.

ANEXO. H ANÁLISIS DE SENSIBILIDAD DEL UMBRAL



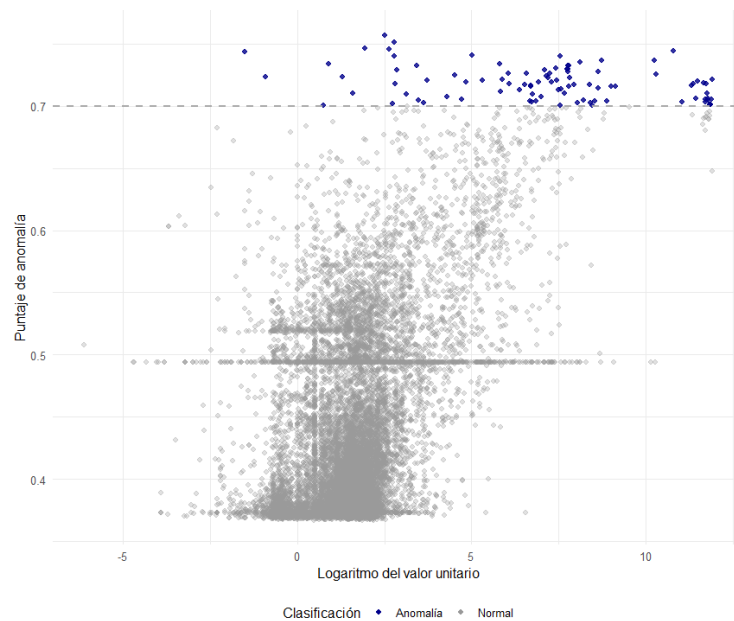
Fuente: Elaboración propia.

ANEXO. I DENSIDAD DEL PUNTAJE DE ANOMALÍA



Fuente: Elaboración propia

ANEXO. J RELACIÓN ENTRE EL PUNTAJE DE ANOMALÍA Y EL LOGARÍTMO DEL VALOR UNITARIO



Fuente: Elaboración propia

ANEXO. K IMPACTO PROMEDIO DE LOS ATRIBUTOS SEGÚN LIME

Atributo	Impacto promedio
$\Delta VU_{m,t}$	0,17900
$\Delta \ln(VU)_t$	0,11100
Z_{cap}	0,03520
Partida sin historia	0,01200
Mes_{cos}	0,00942
Mes_{sin}	0,00755

Fuente: Elaboración propia.

ANEXO. L RESULTADOS DE LA PERMUTACIÓN DE ATRIBUTOS

Atributo permutado	Anomalías recogidas	Variación
Original	94	0 %
$\Delta \ln(VU)_t$	41	-56.38 %
$\Delta VU_{m,t}$	21	-77.66 %
Z_{cap}	79	-15.96 %
Mes_{sin}	94	0 %
Mes_{cos}	94	0 %
Partida sin historia	96	2.13 %

Fuente: Elaboración propia.

ANEXO. M PROCESO DE VALIDACIÓN DE RESULTADOS

Fase	Actividad	Objetivo
I. Selección de anomalías	Filtrado de registros con <i>score</i> > 0,70.	Identificar candidatos a anomalía.
II. Análisis de contexto	Revisión de series históricas del exportador, partida y país.	Descartar <i>shocks</i> de mercado genuinos.
III. Clasificación final	Etiquetado final: Anomalía verdadera vs. falso positivo.	Establecer la métrica de precisión.

Fuente: Elaboración propia.

ANEXO. N EVALUACIÓN DE SENSIBILIDAD MEDIANTE PERTURBACIÓN SINTÉTICA

Realidad del Dato	Predicción del Modelo (EIF)		Total
	<i>Anomalía</i>	<i>Normal</i>	
Anomalía Inyectada	30	20	50
Dato Normal (Control)	0	325	325
Total	30	345	375

Fuente: Elaboración propia.