



Maestría en

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

**Trabajo previo a la obtención de título de Magister en Ciencia de Datos y
Máquinas de Aprendizaje con mención en Inteligencia Artificial**

AUTORES:

Karla Elizabeth Abarca Jaramillo

Diego Fernando Chimbo Yepez

Marcelo Segundo Sevilla Vique

TUTORES:

Andrés Roberto Navas Castellanos

Marcela Azucena Venegas Espinosa

TÍTULO DEL TRABAJO DE TITULACIÓN:

Modelo de detección de patrones anómalos asociados a riesgo de evasión fiscal en
Ecuador.

Quito – Ecuador

Julio - 2026

**REINVENTEMOS
EL FUTURO**

Certificación de Autoría

Nosotros Karla Elizabeth Abarca Jaramillo, Diego Fernando Chimbo Yepez y Marcelo Segundo Sevilla Vique, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

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



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
**KARLA ELIZABETH
ABARCA JARAMILLO**



**Diego Fernando
Chimbo Yepez**

Time Stamping
Security Data

Firma del graduando

Karla Elizabeth Abarca Jaramillo

Firma del graduando

Diego Fernando Chimbo Yepez



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
**MARCELO SEGUNDO
SEVILLA VIQUE**

Firma del graduando

Marcelo Segundo Sevilla Vique

Autorización de Derechos de Propiedad Intelectual

Nosotros, Karla Elizabeth Abarca Jaramillo, Diego Fernando Chimbo Yepez y Marcelo Segundo Sevilla Vique, en calidad de autores del trabajo de investigación titulado *Modelo de detección de patrones anómalos asociados a riesgo de evasión fiscal en Ecuador*, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, (Julio 2026)



Firma del graduando

Karla Elizabeth Abarca Jaramillo



Firma del graduando

Diego Fernando Chimbo Yepez

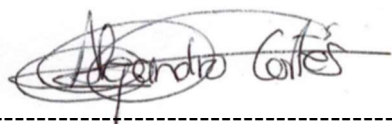


Firma del graduando

Marcelo Segundo Sevilla Vique

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López** Director EIG y **Karla Estefanía Mora Cajas** Coordinadora UIDE, declaramos que: **Karla Elizabeth Abarca Jaramillo, Diego Fernando Chimbo Yopez y Marcelo Segundo Sevilla Vique**, son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López
Director de la
Maestría en Ciencia de datos y máquinas de
aprendizaje con mención en Inteligencia
Artificial



Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Ciencia de datos y máquinas de
aprendizaje con mención en Inteligencia
Artificial

Dedicatoria

A Dios, que ha sido nuestra fuerza, fortaleza y la guía durante este camino.

A nuestras familias, que fueron nuestra principal fuente de motivación y fortaleza durante el desarrollo del trabajo de titulación. Cada esfuerzo y tiempo dedicado ha sido dedicado en ellos, para poder construir un futuro mejor para ellos.

Agradecimientos

A Dios, por la fortaleza brindada para culminar este trabajo de titulación. A todas las personas que han aportado de manera directa o indirecta, contribuyendo al desarrollo de este trabajo, brindando apoyo y guía en las distintas etapas de este proceso académico.

Resumen

La investigación no buscó clasificar contribuyentes individuales ni emitir juicios concluyentes sobre incumplimiento tributario, sino analizar la utilidad de cada modelo para reconocer perfiles agregados de interés analítico según año, provincia y sector económico.

El estudio se desarrolló bajo un enfoque cuantitativo, aplicado y exploratorio, utilizando información de recaudación tributaria, compras y ventas declaradas, tipo de contribuyente y condición de gran contribuyente, correspondiente al periodo 2017-2026, con tratamiento especial de los datos parciales de 2026. La unidad de análisis se definió mediante perfiles agregados, sin identificación de contribuyentes individuales. Se aplicaron procesos de limpieza, homologación, normalización y construcción de variables derivadas, seguidos de técnicas como PCA, UMAP, K-Means, DBSCAN, HDBSCAN, Gaussian Mixture Models e Isolation Forest.

Los resultados permitieron distinguir comportamientos diferenciados y segmentos de alta prioridad analítica, principalmente en provincias y sectores con mayor concentración económica. En consecuencia, los modelos evaluados funcionan como herramientas complementarias para priorizar el análisis fiscal agregado, orientar revisiones posteriores y apoyar futuras investigaciones con datos institucionales de mayor granularidad, sin constituir evidencia definitiva de evasión fiscal.

Palabras Clave: evasión fiscal, aprendizaje no supervisado, detección de anomalías, datos abiertos, riesgo tributario.

Abstract

This study aimed to evaluate how different unsupervised learning models fit the analysis of publicly available aggregated data from the Ecuadorian Internal Revenue Service (SRI), in order to identify anomalous patterns associated with fiscal risk in Ecuador. The research did not seek to classify taxpayers as evaders or issue conclusive judgments on tax noncompliance, but rather to analyze the usefulness of each model for recognizing aggregated profiles of analytical interest by year, province, and economic sector.

The study followed a quantitative, applied, and exploratory approach, using information on tax collection, declared purchases and sales, taxpayer type, and large taxpayer status for the 2017-2026 period, with special treatment for the partial data available for 2026. The unit of analysis was defined through aggregated profiles, without identifying individual taxpayers. Data processing included cleaning, harmonization, normalization, and the construction of derived variables, followed by techniques such as PCA, UMAP, K-Means, DBSCAN, HDBSCAN, Gaussian Mixture Models, and Isolation Forest.

The results made it possible to distinguish differentiated behaviors and high analytical priority segments, mainly in provinces and sectors with greater economic concentration. Therefore, the evaluated models function as complementary tools to prioritize aggregated fiscal analysis, guide further reviews, and support future research using institutional data with greater granularity, without constituting definitive evidence of tax evasion.

Keywords: tax evasion, unsupervised learning, anomaly detection, open data, tax risk

Tabla de Contenidos

Acuerdo de confidencialidad	IV
Resumen.....	VIII
Abstract	IX
CAPÍTULO 1 INTRODUCCIÓN	20
1.1 Título del Proyecto.....	20
1.2 Definición del Proyecto.....	20
1.3 Justificación e importancia del trabajo de investigación.....	24
<i>1.3.1 Justificación Técnica</i>	24
<i>1.3.2 Justificación legal</i>	26
<i>1.3.3 Justificación práctica (económica y operativa).</i>	27
1.4 Naturaleza o tipo de proyecto.....	27
1.5 Alcance del proyecto	29
1.6 Objetivos	30
<i>1.6.1 Objetivo general</i>	30
<i>1.6.2 Objetivos específicos</i>	31
1.7 Metodología	31
CAPÍTULO 2 REVISIÓN DE LITERATURA (MARCO REFERENCIAL)	34
2.1 Estado del Arte.....	34
2.2 Marco Teórico	35
CAPÍTULO 3 GOBERNANZA Y ENTENDIMIENTO DE LOS DATOS	39
3.1 Marco normativo para la reutilización y gobernanza de datos abiertos	39

3.2 Fuentes de información utilizadas.....	40
3.3 Catalogación de diccionarios	42
3.4 Normalización de variables	45
3.5 Alcance analítico de la información disponible	46
3.6 Granularidad y unidad de análisis	47
3.7 Reglas de validación y calidad de datos.....	48
3.8 Síntesis del capítulo.....	49
CAPÍTULO 4 PREPARACIÓN Y LIMPIEZA DE DATOS.....	50
4.1 Fuentes de información utilizada.	51
4.2 Integración y consolidación de información	52
4.2.1 Consolidación histórica de la información	52
4.2.2 Homologación de estructuras.....	52
4.2.3 Entorno tecnológico para preparación de datos.....	53
4.2.4 Limpieza y tratamiento de inconsistencias	54
4.2.4.1 Proceso de limpieza de la información de recaudación tributaria	54
4.2.4.2 Evaluación de calidad de la información de recaudación tributaria	57
4.2.4.3 Proceso de limpieza de la información de compras y ventas declaradas	58
CAPÍTULO 5 CONSTRUCCIÓN DE VARIABLES ANALÍTICAS	63
5.1 Metodología para la construcción de variables analíticas.....	63
5.2 Variables derivadas de la recaudación tributaria.....	65
5.2.1 Recaudación tributaria total anual	65

5.2.2 Variación interanual de la recaudación tributaria	66
5.2.3 Recaudación tributaria total mensual	67
5.2.4 Participación provincial en la recaudación tributaria	69
5.2.5 Participación por grupo de impuesto	70
5.2.6 Participación por tipo de contribuyente.....	72
5.2.7 Participación por condición de gran contribuyente.....	73
5.3 Variables derivadas de compras y ventas declaradas.....	74
5.3.1 Compras y ventas declaradas totales por año.....	75
5.3.2 Variación interanual de compras y ventas declaradas	76
5.3.3 Balance declarado y relación compras/ventas.....	78
5.3.4 Participación provincial en ventas declaradas	79
5.3.5 Participación sectorial en ventas declaradas	81
5.3.6 Comparación sectorial entre ventas y compras declaradas	83
5.3.7 Síntesis de variables derivadas de compras y ventas declaradas	85
CAPÍTULO 6 ANÁLISIS EXPLORATORIO DE DATOS	87
6.1 Análisis exploratorio de la recaudación tributaria.....	87
6.1.1 Concentración territorial de la recaudación tributaria	93
6.1.2 Composición de la recaudación por grupos de impuesto	95
6.1.3 Composición de la recaudación según tipo de contribuyente	96
6.1.4 Composición de la recaudación según condición de gran contribuyente.....	97
6.1.5 Síntesis del análisis exploratorio de la recaudación tributaria	99

6.2 Análisis exploratorio de compras y ventas declaradas	100
CAPÍTULO 7 MODELADO	105
7.1 Construcción de la base de modelado	106
7.2 Construcción del perfil analítico y variables derivadas	107
7.3 Estandarización de variables	108
7.4 Análisis de componentes principales (PCA)	108
7.5 Segmentación mediante K-Means.....	110
7.6 Agrupamiento basado en densidad mediante DBSCAN.....	113
7.7 Agrupamiento jerárquico por densidad mediante HDBSCAN	114
7.8 Modelado probabilístico mediante Gaussian Mixture Models	116
7.9 Detección directa de anomalías mediante Isolation Forest	118
7.10 Visualización no lineal mediante UMAP	119
7.11 Análisis de sensibilidad metodológica	121
7.12 Síntesis del capítulo.....	122
CAPÍTULO 8 PUNTAJE Y PRIORIZACIÓN ANALÍTICA DEL RIESGO	123
8.1 Puntaje de prioridad analítica basado en K-Means	123
8.2 Señales de riesgo tributario	124
8.3 Priorización analítica con K-Means	126
8.4 Priorización basada en modelos de densidad	127
8.5 Priorización basada en incertidumbre probabilística con GMM.....	129
8.6 Priorización mediante Isolation Forest.....	131

8.7 Comparación general de priorización por modelo	132
8.8 Síntesis del capítulo.....	134
CAPÍTULO 9 INTERPRETACIÓN, DISCUSIÓN Y CONCLUSIONES DEL MODELO	135
9.1 Lectura integrada de los modelos aplicados.....	135
9.2 Análisis sectorial de perfiles priorizados	136
9.3 Análisis territorial de perfiles priorizados	138
9.4 Utilidad del puntaje de prioridad analítica como herramienta de apoyo.....	139
9.5 Limitaciones del análisis	140
9.6 Conclusiones del modelado.....	141
9.7 Recomendaciones y trabajo futuro.....	142
Referencias.....	144

Índice de tablas

Tabla 1 Fuentes de información consideradas en la investigación	41
Tabla 2 <i>Variables principales utilizadas en el estudio</i>	43
Tabla 3 <i>Correspondencia entre variables originales y nombres canónicos</i>	44
Tabla 4 <i>Normalización de variables utilizadas en el procesamiento</i>	45
Tabla 5 <i>Fuentes de información utilizadas en la investigación</i>	51
Tabla 6 <i>Resumen de validación inicial de los archivos de recaudación tributaria (2017–2026)</i>	55
Tabla 7 <i>Validación de consistencia estructural y temporal de la información de recaudación tributaria (2017-2026)</i>	56
Tabla 8 <i>Resultados de la validación inicial de calidad de los registros de recaudación tributaria</i>	57
Tabla 9 <i>Resumen de validación estructural de los archivos de compras y ventas declaradas, período 2017–2026</i>	59
Tabla 10 <i>Resultados de la homologación y validación de calidad de compras y ventas declaradas, período 2017–2026</i>	60
Tabla 11 <i>Recaudación tributaria total anual del Ecuador, período 2017–2026</i>	65
Tabla 12 <i>Variación interanual de la recaudación tributaria, período 2017–2025</i>	66
Tabla 13 <i>Participación mensual de la recaudación tributaria por año, período 2017–2025</i>	68
Tabla 14 <i>Participación provincial en la recaudación tributaria, principales provincias, período 2017–2025</i>	69
Tabla 15 <i>Participación por grupo de impuesto en la recaudación tributaria, período 2017–2025</i>	71
Tabla 16 <i>Participación por tipo de contribuyente en la recaudación tributaria, período 2017–2025</i>	72

Tabla 17 <i>Participación por condición normalizada de gran contribuyente, período 2021–2025</i>	73
Tabla 18 <i>Compras y ventas declaradas totales por año, período 2017–2026</i>	75
Tabla 19 <i>Variación interanual de compras y ventas declaradas, período 2017–2025</i>	77
Tabla 20 <i>Balance declarado y ratio compras/ventas, período 2017–2026</i>	78
Tabla 21 <i>Participación provincial en ventas declaradas, principales provincias, período 2017–2025</i>	80
Tabla 22 <i>Participación sectorial en ventas declaradas, principales sectores, período 2017–2025</i>	81
Tabla 23 <i>Comparación sectorial entre ventas y compras declaradas, principales sectores, período 2017–2025</i>	84
Tabla 24 <i>Resumen de construcción de la base de modelado</i>	106
Tabla 25 <i>Variables utilizadas para caracterizar los perfiles tributarios</i>	107
Tabla 26 <i>Varianza explicada por los componentes principales</i>	108
Tabla 27 <i>Coeficiente silhouette para diferentes valores de k</i>	111
Tabla 28 <i>Distribución de perfiles por cluster K-Means</i>	111
Tabla 29 <i>Perfil promedio de los clusters K-Means</i>	111
Tabla 30 <i>Distribución de perfiles por cluster DBSCAN</i>	114
Tabla 31 <i>Distribución de perfiles por cluster HDBSCAN</i>	115
Tabla 32 <i>Distribución de perfiles por cluster GMM</i>	117
Tabla 33 <i>Distribución de perfiles normales y anómalos según Isolation Forest</i>	118
Tabla 34 <i>Distribución de perfiles según puntaje K-Means</i>	124
Tabla 35 <i>Distribución de perfiles según nivel K-Means</i>	124
Tabla 36 <i>Resumen de señales de riesgo tributario identificadas</i>	125
Tabla 37 <i>Distribución del número total de señales por perfil</i>	125

Tabla 38	<i>Distribución de prioridad analítica según K-Means</i>	126
Tabla 39	<i>Principales perfiles priorizados según K-Means</i>	127
Tabla 40	<i>Puntaje, nivel y prioridad según DBSCAN</i>	128
Tabla 41	<i>Principales perfiles priorizados según DBSCAN</i>	128
Tabla 42	<i>Puntaje, nivel y prioridad según HDBSCAN</i>	128
Tabla 43	<i>Principales perfiles priorizados según HDBSCAN</i>	128
Tabla 44	<i>Estadísticos descriptivos del puntaje GMM</i>	129
Tabla 45	<i>Señales y prioridad según GMM</i>	130
Tabla 46	<i>Principales perfiles priorizados según GMM</i>	130
Tabla 47	<i>Puntaje, nivel y prioridad según Isolation Forest</i>	131
Tabla 48	<i>Principales perfiles priorizados según Isolation Forest</i>	131
Tabla 49	<i>Comparación de criterios de priorización por modelo</i>	133
Tabla 50	<i>Síntesis interpretativa de los modelos aplicados</i>	136
Tabla 51	<i>Sectores con mayor cantidad de perfiles de alta prioridad</i>	137
Tabla 52	<i>Provincias con mayor cantidad de perfiles de alta prioridad</i>	138
Tabla 53	<i>Limitaciones principales del proceso de modelado e interpretación</i>	140

Índice de figuras

Figura 1 Fases de pipeline.....	32
Figura 2 Evolución anual de la recaudación tributaria total, período 2017–2025	88
Figura 3 Variación interanual de la recaudación tributaria, período 2018–2025	89
Figura 4 Participación mensual acumulada de la recaudación tributaria, período 2017–2025	90
Figura 5 Distribución mensual relativa de la recaudación tributaria por año, período 2017– 2025.....	92
Figura 6 Participación acumulada de las principales provincias en la recaudación tributaria, período 2017–2025	94
Figura 7 Participación acumulada por grupo de impuesto en la recaudación tributaria, período 2017–2025.....	95
Figura 8 Participación acumulada por tipo de contribuyente en la recaudación tributaria, período 2017–2025	96
Figura 9 Participación acumulada según condición de gran contribuyente, período 2021–2025	98
Figura 10 Evolución anual de compras y ventas declaradas, período 2017–2025	100
Figura 11 Variación interanual de compras y ventas declaradas, período 2018–2025.....	101
Figura 12 Participación acumulada de las principales provincias en ventas declaradas, período 2017–2025.....	102
Figura 13 Participación acumulada por sector económico en ventas declaradas, período 2017– 2025.....	102
Figura 14 Comparación sectorial entre ventas y compras declaradas, período 2017–2025 ..	103
Figura 15 Varianza acumulada del modelo PCA.....	109
Figura 16 Distribución de perfiles tributarios en el espacio PCA	109
Figura 17 Método del codo para selección de clusters en K-Means.....	111

Figura 18 Segmentación K-Means proyectada en el espacio PCA.....	112
Figura 19 Agrupamiento DBSCAN proyectado en el espacio PCA.....	114
Figura 20 Agrupamiento HDBSCAN proyectado en el espacio PCA.....	115
Figura 21 Clusters GMM proyectados en el espacio PCA	117
Figura 22 Anomalías detectadas por Isolation Forest en el espacio PCA	119
Figura 23 Distribución general de perfiles tributarios mediante UMAP.....	120
Figura 24 Clusters K-Means proyectados en el espacio UMAP.....	120
Figura 25 Anomalías de Isolation Forest proyectadas en el espacio UMAP.....	121
Figura 26 Distribución del número de señales de riesgo por perfil	125
Figura 27 Perfiles de alta prioridad identificados por modelo.....	133
Figura 28 Sectores con mayor concentración de perfiles priorizados	137
Figura 29 Provincias con mayor concentración de perfiles priorizados.....	138

CAPÍTULO 1

INTRODUCCIÓN

1.1 Título del Proyecto

Modelo de detección de patrones anómalos asociados a riesgo de evasión fiscal en Ecuador.

1.2 Definición del Proyecto

Previo a la definición del proyecto es indispensable tener en cuenta a qué se refiere la evasión fiscal, según la Real Academia Española se la define como Incumplimiento culpable de la obligación tributaria válidamente surgida de la realización del hecho imponible (RAE, s. f.).

Para León (2014), la evasión fiscal es considerada una conducta ilícita ya que se oculta una parte o todo derecho que está imponible, que realiza el sujeto pasivo, para no pagar la deuda tributaria que está obligado por ley a realizarlo.

Según Aguirre (2019), Se trata de una violación de la ley que establece el deber de pagar al fisco lo que este órgano tiene derecho a exigir, Choez (2019), cuando el contribuyente no cumple con sus obligaciones fiscales, contribuye a la propagación de esta práctica ilícita, lo que genera consecuencias graves para el país al dificultar el cumplimiento de las metas del milenio, relacionadas con alimentación, educación y salud.

Teniendo en cuenta el contexto ecuatoriano, se conceptualiza como el uso de estrategias ilícitas (como la simulación de operaciones o empresas fantasma) por parte de quienes están obligados jurídicamente a pagar, generando un impacto negativo directo en el presupuesto nacional (Vallejo-Vega et al., 2025).

Como se evidencia en las definiciones descritas la evasión fiscal se constituye en un incumplimiento por parte del sujeto pasivo, que es la persona natural o jurídica que, según la

ley, está obligada al cumplimiento de la prestación tributaria, sea como contribuyente o como responsable, la definición de sujeto pasivo se encuentra definida en el Art 24 del código tributario emitido por el SRI (*Codigo_Tributario_20_junio_2023*, s. f.).

Flores et al. (2019), la evasión fiscal es un problema que afecta de forma directa a la capacidad del Estado ecuatoriano para recaudar recursos y sostener el financiamiento de los servicios públicos. Cuando una parte de los ingresos tributarios no se materializa, se generan brechas que impactan en la planificación económica y en la ejecución de políticas públicas, limitando el desarrollo equitativo del país.

Lo antes citado se constituye en una prueba que la evasión fiscal no es solo un problema de “Falta de dinero en caja”, si no en un obstáculo que altera los presupuestos planificados dentro del Plan Nacional de Desarrollo “Ecuador No Se Detiene 2025 - 2029”

Como señala la Comisión Económica para América Latina y el Caribe (CEPAL, 2020), este fenómeno no se presenta de manera homogénea, sino que varía según el tipo de impuesto, la actividad económica y el territorio. Además, puede manifestarse a través de distintos mecanismos, como el uso de empresas interpuestas, estructuras societarias complejas o transacciones entre entidades relacionadas que dificultan la trazabilidad de los flujos económicos, exigiendo el diseño de nuevas metodologías analíticas para su detección.

El Servicio de Rentas Internas pone a disposición información agregada sobre la recaudación tributaria, la cual permite observar su comportamiento a lo largo del tiempo y entre diferentes regiones del país. No obstante, el análisis de estos datos suele limitarse a comparaciones básicas o reportes descriptivos, lo cual dificulta detectar cambios inusuales o desviaciones que se apartan de los patrones habituales de recaudación. En algunos casos, estas variaciones pueden estar relacionadas con riesgos de incumplimiento o con dinámicas atípicas dentro del sistema tributario, incluyendo posibles esquemas de evasión asociados a relaciones entre empresas o a comportamientos inconsistentes con la actividad económica reportada.

El análisis se vuelve aún más complejo debido a que la información disponible es de carácter público (Datos Abiertos) y agregado, por lo que no es posible identificar directamente a contribuyentes individuales ni confirmar la existencia de evasión fiscal, esto debido a la normativa vigente, un claro ejemplo de ello es la Ley Orgánica de Protección de Datos Personales.

En concordancia con lo descrito por el centro Interamericano de Administraciones Tributarias (CIAT) en el ámbito de análisis de datos, los datos permiten observar comportamientos generales que, cuando se desvían de lo esperado, pueden ayudar como señales de alerta para el análisis fiscal. En este sentido, el estudio se enfoca en patrones agregados que podrían estar asociados indirectamente a prácticas como subdeclaración, triangulación de operaciones o estructuras empresariales utilizadas para reducir la carga tributaria (2020_*Manual-gestion-riesgos_CIAT-SII-FMI*, s. f.).

Ante esta situación, resulta pertinente incorporar herramientas de análisis de datos que permitan explorar los datos con mayor profundidad. Como sostienen Chandola et al. (2009), las técnicas de aprendizaje automático no supervisado ofrecen la posibilidad de identificar patrones anómalos sin necesidad de contar con información previa sobre etiquetas institucionales de evasión confirmada. Este enfoque metodológico, validado por el Centro Interamericano de Administraciones Tributarias (CIAT, 2021) para el ámbito fiscal, permite aislar comportamientos atípicos u *outliers* basándose puramente en la estructura interna y las desviaciones intrínsecas de los datos agregados. Por esta razón, el presente estudio se enfoca en el desarrollo de un modelo que permita detectar patrones atípicos en la recaudación tributaria del Ecuador, con el propósito de apoyar el análisis del riesgo de evasión fiscal desde un enfoque basado en datos (2020_*Manual-gestion-riesgos_CIAT-SII-FMI*, s. f.)

El modelo a desarrollar permitirá identificar patrones anómalos asociados al riesgo de evasión fiscal, utilizando datos públicos de recaudación tributaria del Ecuador proporcionados

por el Servicio de Rentas Internas. El enfoque está orientado a detectar desviaciones en el comportamiento agregado de la recaudación que puedan estar vinculadas, de manera indirecta, con prácticas de planificación fiscal agresiva o mecanismos de evasión.

El análisis se basa en información agregada, considerando variables como el período temporal, el tipo de impuesto, la provincia, el cantón y la actividad económica. Mediante la aplicación de técnicas de aprendizaje automático no supervisado, se busca identificar comportamientos atípicos en la recaudación que se desvíen de los patrones históricos observados. Estos patrones pueden reflejar dinámicas inusuales dentro de determinados sectores o territorios, que en algunos casos podrían estar relacionadas con interacciones entre empresas, estructuras societarias o flujos económicos que no siguen comportamientos esperados.

Los resultados obtenidos no constituyen evidencia directa de evasión fiscal, sino señales de riesgo que permiten identificar sectores, regiones o períodos que presentan comportamientos inusuales y que podrían ser objeto de análisis más detallados en estudios posteriores. En este sentido, el modelo propuesto se plantea como una herramienta de apoyo para la priorización del análisis fiscal, y como base para futuras extensiones que incorporen fuentes adicionales de información, como registros empresariales o datos de comercio exterior, con el fin de profundizar en la comprensión de posibles esquemas de evasión.

Para efectos metodológicos, el presente estudio delimita la detección de patrones anómalos al análisis de señales agregadas asociadas a riesgo tributario, sin realizar identificación individual de contribuyentes. La población analítica se conformó por perfiles agregados contruidos a partir de datos públicos del Servicio de Rentas Internas (SRI), organizados por año, provincia y sector económico durante el periodo 2017 al 2026. Las variables analíticas principales consideradas corresponden a recaudación tributaria, compras y ventas declaradas, tipo de contribuyente, condición de gran contribuyente, provincia, sector

económico y medidas derivadas de dispersión, concentración, variación y volumen de registros. En consecuencia, el alcance del modelo es exploratorio y no supervisado; sus resultados deben interpretarse como prioridades analíticas para orientar revisiones posteriores, y no como evidencia concluyente de evasión fiscal.

1.3 Justificación e importancia del trabajo de investigación

Tomando como punto de partida las limitaciones analíticas y las restricciones normativas descritas en la definición del proyecto, la presente investigación se justifica desde las perspectivas técnica, legal y práctica, como se detalla a continuación.

1.3.1 Justificación Técnica

Históricamente, las administraciones tributarias han dependido de sistemas basados en reglas expertas predefinidas que, aun combinados con auditorías aleatorias, presentan limitaciones para identificar oportunamente patrones asociados al riesgo de evasión fiscal. Esto se debe a que los sistemas heurísticos suelen estar diseñados para reconocer tipologías previamente conocidas e incorporadas en reglas lógicas. En la actualidad, la evasión fiscal constituye un fenómeno dinámico y altamente adaptativo, en el que las estrategias utilizadas para eludir controles pueden modificarse de manera continua, lo que reduce la eficacia de enfoques estrictamente estáticos (OECD, 2022).

El empleo de herramientas como Machine Learning supervisado resulta insuficiente debido a la necesidad de contar con datos históricos exhaustivos y etiquetados, a ello se suma que las bases de datos de evasiones confirmadas podrían estar sesgadas hacia quienes ya fueron descubiertos haciendo uso de estas prácticas y métodos antiguos, a ello se suma que se ignoran completamente las nuevas formas de evasión fiscal que aún permanecen ocultas en los datos (Balderas-Díaz & Guerrero-Contreras, s. f.).

En este contexto, la aplicación de técnicas de aprendizaje automático no supervisado que permiten identificar patrones anómalos y comportamientos atípicos en la recaudación, sin necesidad de contar con información etiquetada ni vulnerar la confidencialidad de los contribuyentes. Motivo por el que el presente proyecto propone la implementación de **Machine Learning no supervisado** como la respuesta analítica idónea.

Como destaca Pang, las anomalías representan desviaciones significativas en espacios de datos complejos que no se ajustan a las distribuciones generativas del comportamiento mayoritario (Pang et al., 2021). Desde esta perspectiva, un comportamiento asociado a incumplimiento tributario real podría diferir de manera medible respecto del patrón dominante de contribuyentes o perfiles con características comparables.

La aplicación de algoritmos avanzados de detección de anomalías permite aislar estas observaciones atípicas evaluando la densidad o la topología de los datos. Estos modelos explotan la naturaleza cuantitativa de las anomalías financieras (que son estadísticamente raras y dimensionalmente diferentes) para separarlas de las observaciones normales de manera algorítmica, logrando identificar perfiles de alta prioridad analítica sin necesidad de reglas humanas preconcebidas ni conocimientos previos sobre cómo se ejecutó la evasión.

El grado innovador del modelo propuesto radica en que el modelo se constituye en una herramienta proactiva, preventiva en lugar de los procedimientos reactivos y pasivos actuales. Lo que no solamente optimiza la detección de anomalías en etapas tempranas si no que establece un posible marco tecnológico resiliente, capaz de aprender y adaptarse a la evolución de los diferentes métodos y formas de evasión

Desde una perspectiva académica, el trabajo integra conceptos de ciencia de datos y aprendizaje automático con el análisis tributario, fortaleciendo un enfoque interdisciplinario.

1.3.2 Justificación legal

Al momento de abordar la normativa vigente en el Ecuador existe un amplio bagaje que va desde la Ley de Comercio Electrónico, Firmas y Mensajes de Datos, la Ley Orgánica para la Transformación Digital y Audiovisual, hasta la Ley Orgánica de Protección de Datos Personales, esta última merece especial atención debido a las restricciones que impone en lo referente al procesamiento de los datos personales, así como los derechos ARCO que contempla para el Titular de los mismos (Barrezueta, s. f.).

Las restricciones contempladas en la normativa vigente no constituyen un impedimento para el desarrollo del presente estudio, sino una condición que limita su alcance metodológico. La investigación utiliza exclusivamente datos públicos, abiertos y agregados del Servicio de Rentas Internas (SRI), por lo que no accede a datos puntuales por individuo, datos personales identificables ni información protegida por sigilo tributario. Esta delimitación permite desarrollar un análisis compatible con los principios de protección de datos personales, minimización, finalidad y proporcionalidad.

En este contexto, el modelo no opera sobre personas naturales o jurídicas individualizadas, sino sobre perfiles agregados contruidos por año, provincia y sector económico. Por ello, los resultados no deben interpretarse como actos para determinar evasión tributaria, acusaciones, sanciones ni identificación individual de contribuyentes con incumplimiento tributario. Su propósito es construir una capa de priorización macroanalítica que permita detectar segmentos con comportamientos atípicos y que pueden servir como base para futuras revisiones académicas o institucionales, siempre bajo autorización legal y con fuentes de mayor granularidad y especificidad.

1.3.3 Justificación práctica (económica y operativa).

Una de las barreras más grandes para la transformación digital y la innovación en el sector público ecuatoriano es la incertidumbre tecnológica y los altos costos de implementación de soluciones de software. Bajo esta limitación, la implementación de una prueba de concepto (PoC) o prototipo funcional permite validar de manera empírica la viabilidad del aprendizaje no supervisado para identificar patrones anómalos asociados a riesgo de evasión antes de comprometer presupuesto institucional, proporcionando evidencia medible sobre la utilidad del modelo (Zuiderwijk et al., 2021).

Al no depender de costosa infraestructura de Big Data, el prototipo puede ser entrenado y testado empleando entornos de desarrollo de código abierto (Python y sus librerías de aprendizaje automático) (*Consensos y conflictos en la política tributaria de América Latina | CEPAL*, s. f.), sobre los conjuntos de datos agregados publicados por el SRI.

El principal impacto práctico del presente proyecto radica en aportar una alternativa analítica frente a la escasez de talento humano y recursos técnicos disponibles para revisar grandes volúmenes de información tributaria. Las revisiones fiscales requieren personal calificado, tiempo y criterios técnicos especializados. Por ello, el modelo propuesto no sustituye las facultades institucionales del Servicio de Rentas Internas (SRI), sino que ofrece una capa exploratoria de priorización basada en datos para orientar la atención hacia segmentos agregados con mayor interés analítico (Yarygina, 2021).

1.4 Naturaleza o tipo de proyecto

El presente proyecto de investigación se define por su naturaleza como un **proyecto de investigación aplicativo y tecnológico**, sustentado en un enfoque cuantitativo y un diseño exploratorio.

Aplicativo, debido a que se orienta a la implementación práctica de técnicas de aprendizaje automático para el análisis de la problemática de evasión fiscal en el Ecuador. En tal virtud, el proyecto busca trasladar herramientas de análisis de datos a un contexto aplicado, donde la implementación de un modelo de aprendizaje no supervisado pueda contribuir a la detección de patrones anómalos en el ámbito de la evasión fiscal (Matriano, 2024).

Desde el punto de vista metodológico, el estudio adopta un enfoque cuantitativo, ya que se basa en el análisis de datos numéricos de recaudación tributaria. Asimismo, presenta un carácter exploratorio, puesto que no parte de hipótesis previamente definidas, sino que busca identificar patrones y anomalías a partir de la información disponible. Este enfoque permite analizar comportamientos agregados y detectar desviaciones que podrían estar asociadas a dinámicas atípicas dentro del sistema tributario.

El empleo de técnicas de aprendizaje no supervisado resulta pertinente porque los datos utilizados no cuentan con etiquetas institucionales que permitan confirmar evasión fiscal, lo que impide el uso de modelos de aprendizaje supervisado tradicional (Delua, 2024). En este contexto, el modelo se orienta a la detección de patrones inusuales en los datos proporcionados, los cuales pueden interpretarse como señales de interés analítico a nivel agregado.

Debido a la alta dimensionalidad de las matrices de datos de recaudación tributaria, es necesario el emplear modelos de reducción del espacio geométrico con la finalidad de mitigar la denominada “maldición de la dimensionalidad”, empleando PCA Y UMAP.

En el ámbito de los algoritmos de agrupamiento (Clustering), cuyo objetivo es estructurar el universo de datos en clústeres de comportamiento similar, identificando ciertas observaciones que se alejan de los centroides, en el presente proyecto se emplean K-Means, DBSCAN, HDBSCAN, GMM (Delua, 2024).

Debido a que el presente trabajo se centra en la detección de anomalías no se puede dejar de probar la implementación del modelo empleando Isolation Forest, ya que este modelo

no supervisado basado en árboles de decisión aleatorios está diseñado específicamente para el aislamiento directo de anomalías financieras sin requerir el perfilado previo de los datos normales (González, s. f.). Su eficiencia radica en que las observaciones anómalas requieren significativamente menos particiones estructurales para ser aisladas, convirtiéndose en el estándar analítico para auditar volúmenes masivos de transacciones agregadas.

A partir de lo expuesto, se evidencia que la investigación tiene como propósito evaluar de manera comparativa la utilidad de diversos modelos no supervisados, con el fin de identificar una arquitectura analítica adecuada para reconocer patrones anómalos asociados a riesgo tributario dentro del alcance exploratorio del estudio.

1.5 Alcance del proyecto

El presente proyecto tiene como alcance el diseño y aplicación de un enfoque exploratorio basado en modelos de aprendizaje no supervisado, orientado a identificar perfiles agregados con comportamientos atípicos asociados a riesgo tributario en Ecuador. La unidad de análisis corresponde a combinaciones año-provincia-sector económico construidas a partir de datos públicos agregados del Servicio de Rentas Internas (SRI). Por tanto, los modelos aplicados no identifican contribuyentes individuales, no determinan responsabilidades tributarias y no constituyen evidencia legal de evasión fiscal. Su utilidad se limita a priorizar segmentos que requieren revisión analítica complementaria.

El análisis se restringe únicamente a personas jurídicas debido a que el sector corporativo presenta una mayor consistencia, homogeneidad y estandarización en los registros contables y declaraciones sectoriales lo que permite optimizar la comparabilidad y la precisión de los modelos de agrupamiento.

El modelo se nutrirá únicamente de los conjuntos de datos abiertos de acceso público provistos por el Servicio de Rentas Internas (SRI).

Ante la ausencia de datos etiquetados que permitan confirmar evasión fiscal, el estudio propone una aproximación probabilística basada en asimetrías financieras y desviaciones estadísticas. Los algoritmos de reducción de dimensionalidad, agrupamiento y detección de valores atípicos operan sobre criterios medibles como diferencias entre ventas y recaudación, variaciones temporales atípicas e irregularidades en el comportamiento agregado de los registros sectoriales.

Se considera necesario especificar las exclusiones del estudio: el presente proyecto no identifica contribuyentes individualizados, no constituye prueba legal de evasión fiscal, no confirma fraude y no sustituye las facultades del SRI como ente rector y regulador de las rentas en el Ecuador. Su finalidad es servir como una herramienta de apoyo analítico basada en datos agregados.

La aplicación del presente será válida únicamente en el contexto de estudio y en el territorio nacional ecuatoriano, teniendo en cuenta que la base de la analítica son los datos proporcionados por el Servicio de Rentas Internas SRI, en el formato de datos abiertos.

Si bien el proyecto actual se limita a los datos proporcionados por el SRI, la arquitectura y el pipeline se estructuran de forma modular, lo que posibilita la compatibilidad técnica para incorporarlo en fases futuras con datos de la Superintendencia de Compañías.

1.6 Objetivos

1.6.1 Objetivo general

Aplicar y comparar modelos de aprendizaje no supervisado, con técnicas de agrupamiento, agrupamiento basado en densidad, modelado probabilístico y detección directa de anomalías, para identificar patrones atípicos en el comportamiento agregado de la recaudación tributaria, compras y ventas declaradas asociadas a riesgo fiscal en Ecuador, utilizando datos públicos del Servicio de Rentas Internas (SRI) durante el periodo 2017–2026.

1.6.2 *Objetivos específicos*

- Analizar estadísticamente el comportamiento histórico de la recaudación tributaria, compras y ventas declaradas, provincia, sector económico, tipo de contribuyente y condición de gran contribuyente, a partir de datos públicos del Servicio de Rentas Internas (SRI).
- Construir perfiles tributarios agregados por año, provincia y sector económico, incorporando variables derivadas de volumen, dispersión, concentración, variación y comportamiento relativo de los registros analizados.
- Aplicar técnicas de reducción de dimensionalidad, agrupamiento y detección de anomalías, incluyendo PCA, UMAP, K-Means, DBSCAN, HDBSCAN, Gaussian Mixture Models e Isolation Forest, para identificar segmentos con comportamientos atípicos.
- Comparar la utilidad analítica de cada modelo aplicado mediante criterios de interpretabilidad, consistencia de los perfiles priorizados y capacidad para generar señales de interés tributario agregado.
- Interpretar los patrones identificados mediante un puntaje de prioridad analítica y señales estadísticas complementarias, delimitando los resultados como apoyo al análisis fiscal agregado y no como evidencia concluyente de evasión fiscal.

1.7 Metodología

La presente investigación emplea un enfoque metodológico cuantitativo, de tipo empírico-analítico, estructurado bajo el estándar CRISP-DM (Cross Industry Standard Process for Data Mining). El ciclo de vida del proyecto se operacionaliza mediante un proceso estructurado de ciencia de datos orientado al análisis de patrones atípicos en datos públicos del SRI organizados por año, provincia y sector económico, sin identificar contribuyentes

individuales. Debido a la ausencia de etiquetas institucionales que permitan confirmar evasión fiscal, el estudio adopta un enfoque de aprendizaje no supervisado, enfocado en la segmentación de perfiles, la detección de comportamientos anómalos o desviados del patrón común y la priorización analítica de segmentos de interés fiscal.

A nivel arquitectónico, el flujo metodológico se despliega como un pipeline analítico y secuencial y modular que ha sido diseñado para de esta manera garantizar la reproducibilidad y escalabilidad del modelo en el sector público ecuatoriano, a saber, el pipeline se compone de 5 fases interconectadas que se muestran en el siguiente gráfico:

Figura 1 Fases de pipeline



Ingreso/Extracción de datos: hace referencia la recopilación de los conjuntos de datos abiertos y registros públicos que provee el SRI, se diseña bajo un enfoque de arquitectura abierta con la finalidad de ser compatible para integraciones futuras con data de otras entidades.

Preprocesamiento: los datos crudos se someten a un proceso de extracción, transformación y carga, esta fase contempla la limpieza de registros nulos, la imputación estadística de valores faltantes, la eliminación de redundancias además de la estandarización escalar de variables numéricas para neutralizar las distorsiones existentes.

Análisis exploratorio de datos: en esta etapa se caracteriza la distribución estadística y se identifica las propiedades geométricas del espacio de características, se calculan correlaciones sectoriales además de las asimetrías entre los niveles de ventas, recaudación real y la carga impositiva efectiva, estableciendo las hipótesis iniciales de comportamiento atípico a nivel macro-agregado.

Modelo no supervisado: esta fase constituye el elemento central del presente proyecto. Ante la ausencia de una variable objetivo que permita confirmar evasión fiscal, se aplican

técnicas de reducción de dimensionalidad, agrupamiento, agrupamiento basado en densidad, modelado probabilístico y detección directa de anomalías. K-Means permite segmentar perfiles de comportamiento similar; DBSCAN y HDBSCAN identifican agrupamientos según densidad y perfiles tipo ruido; GMM permite analizar la incertidumbre de pertenencia a un grupo; e Isolation Forest prioriza observaciones que se aíslan con mayor facilidad respecto del comportamiento dominante.

Puntaje y evaluación: una vez aplicados los modelos no supervisados, los resultados se integran mediante un puntaje de prioridad analítica. Este puntaje sintetiza señales estadísticas asociadas a volumen, dispersión, concentración, variación y comportamiento relativo de los perfiles analizados. El objetivo que persigue no es estimar una probabilidad real de evasión fiscal ni automatizar auditorías, sino ordenar segmentos según su interés analítico para orientar de mejor manera los segmentos que necesitan ser analizados con más detalle a posterior, comparaciones entre modelos y posibles investigaciones con fuentes de mayor detalle.

CAPÍTULO 2

REVISIÓN DE LITERATURA (MARCO REFERENCIAL)

2.1 Estado del Arte

Tradicionalmente, las administraciones tributarias dependen de auditorías aleatorias o sistemas heurísticos basados en reglas fijas definidas por expertos. En la actualidad, el estudio de patrones asociados a evasión fiscal puede apoyarse en la ciencia de datos; no obstante, la literatura contemporánea demuestra que los comportamientos de incumplimiento evolucionan con rapidez, motivo por el cual el aprendizaje automático se ha posicionado como una herramienta de optimización para la fiscalización proactiva.

A nivel internacional, (Hernández et al., 2025) desarrollaron un marco analítico empleando el aprendizaje no supervisado para la identificación de anomalías financieras complejas. Su investigación demostró que los enfoques no supervisados poseen una ventaja crítica sobre los supervisados en los entornos gubernamentales: la capacidad de descubrir patrones atípicos no documentados superando de esta manera el sesgo que afecta a los modelos entrenados únicamente con registros previamente detectados.

Empleando un enfoque similar se abordaron la integración de la ciencia de datos en la administración pública, demostrando que el uso de pipelines analíticos, modulares incrementa la precisión en la asignación de puntajes de riesgo en los sectores corporativos, optimizando el despliegue del escaso contingente humano de auditores (Judith et al., 2023).

En el ámbito metodológico de los algoritmos no supervisados, la investigación se ha centrado en la resiliencia de los modelos basados en densidad frente a los modelos basados en árboles de aislamiento, (Radosław J. WOŹNIAK, 2025) realizó un análisis comparativo entre técnicas de agrupamiento (*clustering*) tradicionales y métodos dedicados a la detección de *outliers*. Sus hallazgos revelaron que, si bien algoritmos como *K-Means* son eficientes para

establecer perfiles base de comportamiento económico normal por sectores, los modelos basados en densidad variable, como *HDBSCAN*, poseen una sensibilidad superior para aislar anomalías en conjuntos de datos financieros altamente asimétricos y no lineales.

A nivel regional y local en América latina, la literatura destaca las restricciones impuestas por la disponibilidad de datos de alta granularidad y las normativas de privacidad. (Mohd Padil et al., 2021), en su revisión sistemática sobre el incumplimiento fiscal mediante analítica avanzada, señalaron que la tendencia global se orienta al uso de modelos preservadores de la privacidad. Estos modelos operan transformando las declaraciones nominales de los contribuyentes en vectores numéricos de características económicas abstractas (como ratios de liquidez o proporciones de deducción sectorial). Este enfoque fue validado en el contexto andino por (Hernández et al., 2025) quienes implementaron una arquitectura híbrida de *clustering* jerárquico y *One-Class SVM* para identificar anomalías en transacciones comerciales sin vulnerar la confidencialidad de los datos de origen, demostrando que los datos abiertos agregados poseen la riqueza estadística suficiente para encender alertas tempranas de riesgo fiscal.

2.2 Marco Teórico

Evasión Fiscal y el análisis de riesgo tributario: teniendo en cuenta la perspectiva de la política fiscal contemporánea, la evasión tributaria se define como la acción ilícita y consciente orientada a evitar total o parcialmente el pago de las obligaciones impositivas mediante la ocultación de ingresos o la simulación de deducciones. A diferencia de la elusión fiscal (que explota vacíos legales), la evasión violenta directamente la normativa vigente.

El gestionar el fenómeno de la evasión fiscal empleando un enfoque basado en riesgos (EBR) permite a las agencias de recaudación no asumir una fiscalización universal u homogénea, sino que segmentan a los sujetos pasivos mediante una función de riesgo matemático:

$$\textbf{Riesgo} = \textbf{Probabilidad de incumplimiento} * \textbf{Impacto Económico}$$

El objetivo de la ciencia de datos en el presente ámbito es calcular la manera algorítmica dicha probabilidad a partir del comportamiento transaccional del sujeto pasivo.

Paradigma del aprendizaje automático no supervisado: en el ámbito de la inteligencia artificial, el aprendizaje automático no supervisado se caracteriza por procesar un conjunto de datos de entrada, que carece de etiquetas o variables objetivo-preclasificadas. En la detección de anomalías fiscales este paradigma opera bajo la premisa económica de que la gran mayoría de las empresas operan dentro de los márgenes de la normalidad legal, provocando que las conductas evasivas se manifiesten como observaciones escasas, aisladas y estadísticamente raras (*outliers*).

Reducción de dimensionalidad: La reducción de la dimensionalidad es un método para representar un conjunto de datos utilizando un menor número de características (es decir, dimensiones) sin perder las propiedades significativas de los datos originales. Esto equivale a eliminar características irrelevantes o redundantes, o simplemente datos ruidosos, para crear un modelo con un menor número de variables (Kavlakoglu, 2024).

Análisis de Componentes Principales (PCA): Es una transformación lineal ortogonal que proyecta los datos a un nuevo sistema de coordenadas de menor dimensión. Matemáticamente, busca maximizar la varianza del primer componente principal, garantizando que la pérdida de información sea mínima al reducir variables correlacionadas (International Encyclopedia of Statistical Science, 2025).

Aproximación y Proyección de Colector Uniforme (UMAP): Es una técnica de reducción de dimensionalidad no lineal fundamentada en la geometría riemanniana y la topología algebraica. Asume que los datos financieros se distribuyen a lo largo de un colector (*manifold*) localmente conectado; UMAP preserva tanto la estructura local de los clústeres

como la relación global entre ellos, permitiendo visualizar la separación de perfiles atípicos de forma óptima (International Encyclopedia of Statistical Science, 2025).

Taxonomía de los algoritmos de Agrupamiento (Clustering): el agrupamiento consiste en particionar el conjunto de datos en subclases homogéneas.

Modelos basados en centroides K-Means: Algoritmo iterativo que particiona los datos en K clústeres definidos por sus centroides (μ_K). El modelo minimiza la función de costo de la suma de errores cuadráticos intraclúster (WCSS):

$$J = \sum_{K=1}^K \sum_{X_i \in C_K} ||X_i - \mu_K||^2$$

Establece la línea base de los perfiles normales de recaudación sectorial (Dabbura, 2022).

Modelos basados en densidad (DBSCAN y HDBSCAN): *DBSCAN* define clústeres como regiones continuas de alta densidad de puntos separadas por regiones de baja densidad. Clasifica explícitamente como "ruido" (anomalías) a los puntos que no cumplen con el umbral de densidad mínima (*MinPts*) en un radio ϵ (Chen Jinyin & Yang Dongyong, 2018). *HDBSCAN* extiende este principio de forma jerárquica, permitiendo descubrir clústeres de densidades variables en espacios no lineales, lo cual es crítico para comparar provincias con diferentes escalas de recaudación, entre las ventajas de HDBSCAN se encuentran la sensibilidad a densidad de variables, manejo explícito del ruido, puntuación de anomalías nativa (Campello et al., 2013).

Modelos Probabilísticos (Gaussian Mixture Models - GMM): Asumen que el universo de contribuyentes está compuesto por una superposición de múltiples distribuciones gaussianas independientes. Utiliza el algoritmo de Esperanza-Maximización (EM) para asignar a cada empresa una probabilidad blanda (*soft clustering*) de pertenencia a un grupo, permitiendo modelar transiciones de riesgo no absolutas (Noble, 2025a).

Aislamiento Directo de Anomalías (Isolation Forest): a diferencia de los modelos de clustering que infieren las anomalías de manera indirecta (como subproductos de los grupos normales), *Isolation Forest* aísla las observaciones anómalas de forma directa y explícita. El algoritmo construye un conjunto de árboles de decisión aleatorios (*iTrees*). Dado que las anomalías financieras poseen características transaccionales atípicas y extremas, requieren significativamente menos particiones o "cortes" aleatorios para quedar completamente aisladas en las ramas terminales del árbol. Matemáticamente, la longitud de la trayectoria $h(x)$ de una anomalía en el árbol es notablemente más corta que la de un punto normal, lo que permite calcular un puntaje de anomalía eficiente y escalable para grandes volúmenes de datos agregados (Dhanesha & Mehta, 2025).

CAPÍTULO 3

GOBERNANZA Y ENTENDIMIENTO DE LOS DATOS

La gobernanza de datos comprende el conjunto de políticas, procedimientos, estándares y controles que permiten garantizar la calidad, consistencia, trazabilidad y uso adecuado de la información empleada durante la investigación.

En la presente investigación, la gobernanza se materializa mediante la catalogación de diccionarios de datos, la normalización de variables, la definición de niveles de granularidad, la aplicación de reglas de validación y calidad, así como el cumplimiento de las disposiciones legales aplicables al tratamiento y reutilización de información pública de carácter tributario.

Debido a que los resultados obtenidos dependen directamente de la calidad y comprensión de los datos utilizados, resulta necesario analizar previamente las características de las fuentes de información, la estructura de los datasets, el significado de sus variables, sus limitaciones y las reglas de calidad que regulan su utilización. Esta etapa permite construir una base analítica consistente y confiable para el desarrollo de los procesos posteriores de preparación de datos, análisis exploratorio y detección de patrones asociados al riesgo tributario.

3.1 Marco normativo para la reutilización y gobernanza de datos abiertos

La presente investigación utiliza información pública publicada por el Servicio de Rentas Internas (SRI) dentro del ecosistema nacional de datos abiertos del Ecuador. Este uso se encuentra respaldado por la Política de Datos Abiertos emitida mediante el Acuerdo Ministerial MINTEL-MINTEL-2022-0021, cuyo objetivo es fortalecer la transparencia gubernamental, promover la participación ciudadana, impulsar la investigación científica y fomentar la innovación mediante el uso de información pública (MINTEL, 2022).

Dicha normativa establece que los datos abiertos deben cumplir principios fundamentales como accesibilidad, oportunidad, interoperabilidad, reutilización, comparabilidad y calidad. En este contexto, los principios de interoperabilidad y comparabilidad sustentan la necesidad de implementar procesos de homologación de estructuras, normalización de variables y estandarización de formatos entre los diferentes conjuntos de datos empleados en esta investigación, garantizando así su integración y análisis de manera consistente y confiable (MINTEL, 2022).

Asimismo, la política dispone que la reutilización de información pública debe observar las regulaciones vigentes relacionadas con la protección de datos personales, el acceso a la información pública y las restricciones legales aplicables a información clasificada o reservada (MINTEL, 2022). En concordancia con estas disposiciones, la presente investigación emplea exclusivamente información agregada y de acceso público, sin incorporar datos individuales sujetos a reserva tributaria, garantizando el cumplimiento de los principios éticos y legales aplicables al tratamiento de datos.

3.2 Fuentes de información utilizadas

La presente investigación utiliza diversos conjuntos de datos publicados por el Servicio de Rentas Internas (SRI), seleccionados por su disponibilidad, cobertura histórica y capacidad para representar diferentes dimensiones del comportamiento tributario y económico del país. Entre estas fuentes, el dataset de recaudación tributaria constituye la principal base de análisis debido a su cobertura temporal desde 2017, periodicidad mensual y nivel de detalle disponible para el análisis territorial, sectorial y tributario.

Como fuentes de información complementarias se dispone de datasets relacionados con ventas, contribuyentes activos, inscripciones, cierres, Registro Único de Contribuyentes (RUC), agregadores de pago, mercados en línea, presión fiscal, Índice de Actividad Económica

No Petrolera (IAENP) y catastro de empresas fantasma. Las principales fuentes consideradas se presentan en la Tabla 1.

Tabla 1

Fuentes de información consideradas en la investigación

Dataset	Descripción	Periodicidad	Uso dentro de la investigación
Recaudación de impuestos	Valores recaudados por impuesto, provincia, actividad económica y tipo de contribuyente	Mensual (2017– actualidad)	Fuente principal para identificación de patrones tributarios
Ventas y compras	Información agregada derivada de declaraciones tributarias	Periódica	Contraste de actividad económica
RUC	Contiene información pública del Registro Único de Contribuyentes por Provincia	Actualización institucional	Identificación y clasificación de contribuyentes
Contribuyentes Activos	Registro de contribuyentes activos	Anual	Caracterización del universo tributario
Inscripciones	Nuevos contribuyentes registrados	Anual	Análisis de apertura de actividades económicas
Cierres	Contribuyentes que cesan actividades	Anual	Análisis de dinámica empresarial
Agregadores de Pago	Contiene información de los contribuyentes calificados como agregadores de pago	Periódica	Contexto de actividad económica digital
Mercados en Línea	Contiene información de los contribuyentes calificados como Mercados en línea	Periódica	Complemento de actividad económica
Listado de empresas fantasmas y personas naturales con transacciones inexistentes	Contiene información pública de los contribuyentes que para efectos tributarios son considerados como empresas fantasmas, sociedades o personas naturales con actividades supuestas y/o transacciones inexistentes	Actualización institucional	Referencia contextual de riesgo
Índice de actividad económica no petrolera	Corresponde al índice trimestral total de los cuatro sectores económicos	Mensual	Variable macroeconómica de contexto
Presión Fiscal	Contribución que realizan las empresas y las personas naturales (físicas) al Estado como porcentaje del PIB	Anual	Indicador económico complementario

Nota: Elaboración propia a partir de información pública del Servicio de Rentas Internas (SRI) y de fuentes consideradas para el entendimiento de datos de la investigación.

3.3 Catalogación de diccionarios

Cada dataset cuenta con un diccionario oficial proporcionado por el SRI, el cual documenta el significado, formato y características de las variables disponibles.

Los diccionarios oficiales presentan una estructura estandarizada que incluye información referente al código de la variable, nombre, definición conceptual, tipo de variable, tipo de dato, formato, longitud, categorías permitidas, reglas de validación, así como aspectos relacionados con la confidencialidad, integridad, disponibilidad y clasificación de la información. Esta documentación constituye la base para la correcta interpretación y utilización de los datos dentro de la investigación.

La catalogación de diccionarios constituye una etapa fundamental dentro del entendimiento del dato, ya que permite identificar el significado de las variables, sus formatos, restricciones y posibles relaciones con otras fuentes de información. Esta actividad facilita la correcta interpretación de los datos y proporciona una base consistente para los procesos posteriores de normalización, integración y construcción de variables analíticas. Asimismo, permite asegurar que las transformaciones realizadas durante la preparación de datos conserven el significado original de la información publicada por el SRI, fortaleciendo la trazabilidad y reproducibilidad de la investigación.

Tabla 2*Variables principales utilizadas en el estudio*

Variable	Dataset principal	Descripción
ANIO	Recaudación	Se refiere al año en el que fueron recaudados los valores.
MES	Recaudación	Se refiere al mes en el que fueron recaudados los valores.
PROVINCIA	Recaudación	División político administrativa conformada por la unión de dos o más cantones en donde fue recaudado el tributo. Adicional, la categoría "No tiene" corresponde a lo recaudado de contribuyentes que no tienen asignada una ubicación geográfica específica, puesto que, no están registrados en el RUC.
CANTON	Recaudación	División político administrativa conformada por la unión de dos o más parroquias en donde fue recaudado el tributo. Adicional, la categoría "No tiene" corresponde a lo recaudado de contribuyentes que no tienen asignada una ubicación geográfica específica, puesto que, no están registrados en el RUC.
TIPO_CONTRIBUYENTE	Recaudación / RUC	Clasificación agregada del contribuyente, pueden ser: Naturales y Sociedades. Por otro lado, para los valores recaudados por individuos que no están registrados en el RUC, se adiciona la categoría "No tiene".
GRAN_CONTRIBUYENTE	Recaudación	Presenta información de recaudación a nivel de grandes contribuyentes. Se clasifican en: Q los administrados por la Zonal 9 (Pichincha); G los administrados por la Zonal 8 (Guayas); S los administrados por la Dirección Nacional (Nacional); y, NO los que no son Grandes contribuyentes.
GRUPO_IMPUESTO	Recaudación	Contiene la clasificación de los tributos que administra el SRI a Nivel 1.
CODIGO_OPERA_FAMILIA	Recaudación	Contiene la clasificación de los sectores de actividad económica según CIU 4.0 Nivel 1.
IMPUESTO	Recaudación	Tipo de impuesto administrado por el SRI
SUBGRUPO_IMPUESTO	Recaudación	Contiene la clasificación de los tributos que administra el SRI a Nivel 2.
VALOR_RECAUDADO	Recaudación	Valor recaudado del período económico
ACTIVIDAD_ECONOMICA	RUC	Se refiere a la descripción de la actividad económica registrada por el sujeto pasivo
ESTADO_CONTRIBUYENTE	RUC	Se refiere al estado del registro del sujeto pasivo

Nota. Elaboración propia a partir de los diccionarios oficiales publicados por el Servicio de Rentas Internas

(SRI).

Tabla 3
Correspondencia entre variables originales y nombres canónicos

Campo del diccionario	Campo esperado del Dataset	Tipo	Reglas de limpieza	Riesgo de privacidad operativo
ANIO	ANIO	Entero	coerción a int; rango 2017–actual	Bajo
MES	MES	texto categórico	trim; mapear a mes_num y mes_nombre; validar 12 valores	Bajo
GRUPO_IMPUESTO	GRUPO_IMPUESTO	Texto	mayúsculas, sin espacios dobles	Bajo
SUBGRUPO_IMPUESTO	SUBGRUPO_IMPUESTO	Texto	mayúsculas, trim	Bajo
IMPUESTO	IMPUESTO	Texto	mayúsculas, trim	Bajo
GRAN_CONTRIBUYENTE	GRAN_CONTRIBUYENTE	categórico/binario	normalizar SI/NO; si constante, excluir del modelo	Bajo
CODIGO_OPERA_FAMILIA	CODIGO_OPERA_FAMILIA	Categórico	trim; validar letras CIU y excepciones como NO TIENE/9	Bajo a medio en celdas muy escasas
TIPO_CONTRIBUYENTE	TIPO_CONTRIBUYENTE	Categórico	normalizar dominios; separar SOCIEDADES para v1	Bajo
PROVINCIA	PROVINCIA	Texto	mayúsculas; homologar tildes y variantes	Bajo
CANTON	CANTON	Texto	mayúsculas; homologar tildes y variantes	Bajo a medio en cruces muy escasos
VALOR_RECAUDADO	VALOR_RECAUDADO	numérico decimal	reemplazar coma decimal; convertir a float; validar no negativos	Bajo
ACTIVIDAD_ECONOMICA	ACTIVIDAD_ECONOMICA	Texto	trim; normalización de caracteres especiales; homologación de formato de texto	Bajo

Campo del diccionario	Campo esperado del Dataset	Tipo	Reglas de limpieza	Riesgo de privacidad operativo
ESTADO_CONTRIBUYENTE	ESTADO_CONTRIBUYENTE	texto categórico	normalización de categorías según catálogo oficial del RUC; eliminación de espacios innecesarios	Bajo

Nota. Elaboración propia a partir de la revisión de diccionarios de datos, campos esperados en los datasets y criterios de limpieza definidos para la investigación.

3.4 Normalización de variables

La normalización de variables tiene como finalidad garantizar la interoperabilidad entre los distintos conjuntos de datos utilizados en la investigación. Este proceso permite reducir inconsistencias derivadas de diferencias de nomenclatura, formatos de almacenamiento y criterios de clasificación, facilitando la integración de información proveniente de múltiples fuentes y períodos.

Entre las principales acciones de normalización se consideran:

- Homologación de nombres de variables y de categorías económicas
- Estandarización de formatos, fecha y tipos de contribuyente.
- Normalización de provincias y cantones.

La Tabla 4 presenta la correspondencia entre las variables originales y los nombres canónicos utilizados durante el procesamiento.

Tabla 4
Normalización de variables utilizadas en el procesamiento

Variable original	Nombre canónico
ANIO	anio
MES	mes_label
GRUPO_IMPUESTO	grupo_impuesto
SUBGRUPO_IMPUESTO	subgrupo_impuesto
IMPUESTO	impuesto

Variable original	Nombre canónico
GRAN_CONTRIBUYENTE	gran_contribuyente_flag
CODIGO_OPERA_FAMILIA	ciiu_familia
TIPO_CONTRIBUYENTE	tipo_contribuyente
PROVINCIA	provincia
CANTON	canton
VALOR_RECAUDADO	valor_recaudado

Nota. Elaboración propia a partir del proceso de homologación de nombres de variables aplicado durante el entendimiento y preparación de datos.

3.5 Alcance analítico de la información disponible

El enfoque metodológico de la investigación no busca identificar contribuyentes individuales ni establecer mecanismos de fiscalización individual. Su propósito consiste en detectar patrones sectoriales, territoriales y temporales que permitan construir señales de comportamiento tributario agregado a partir de información pública. Estas señales podrán utilizarse posteriormente como insumo para procesos de priorización analítica o estudios complementarios que integren información adicional proveniente del Registro Único de Contribuyentes (RUC) u otras fuentes públicas disponibles.

A partir de los datos disponibles, es posible realizar los siguientes análisis:

- Analizar la recaudación por provincia.
- Analizar la recaudación por actividad económica.
- Comparar tipos de contribuyentes.
- Construir perfiles tributarios.
- Aplicar técnicas de agrupamiento.
- Identificar patrones atípicos mediante detección de anomalías.
- Generar indicadores de comportamiento tributario agregado.

Por lo tanto, los resultados obtenidos deben interpretarse como señales de comportamiento tributario agregado y no como evidencia concluyente de evasión fiscal.

3.6 Granularidad y unidad de análisis

La información publicada por el SRI posee una estructura agregada y no contiene registros individuales de contribuyentes. Por ello, la unidad de análisis se define a partir de combinaciones de variables temporales, territoriales, económicas y tributarias.

Para efectos de esta investigación, la unidad de análisis corresponde a combinaciones agregadas de año, mes, provincia, cantón, actividad económica, tipo de contribuyente e impuesto. Esta estructura permite generar perfiles comparables entre sectores y territorios, manteniendo el análisis dentro de los límites establecidos por la disponibilidad de datos abiertos.

Entre las dimensiones disponibles se encuentran:

- Año.
- Mes.
- Provincia.
- Cantón.
- Actividad económica (CIIU).
- Tipo de contribuyente.
- Tipo de impuesto.

De acuerdo con la estructura del dataset de recaudación, cada observación se encuentra definida por la combinación de las variables ANIO, MES, GRUPO_IMPUESTO, SUBGRUPO_IMPUESTO, IMPUESTO, GRAN_CONTRIBUYENTE, CODIGO_OPERA_FAMILIA, TIPO_CONTRIBUYENTE, PROVINCIA y CANTON. Esta combinación constituye la llave de agregación principal utilizada para la construcción de perfiles analíticos y estructuras comparables dentro de la investigación. Esta granularidad permite desarrollar análisis de segmentación, clustering y detección de anomalías sobre grupos

homogéneos de información; sin embargo, no permite identificar contribuyentes individuales ni determinar la existencia de evasión fiscal individual.

3.7 Reglas de validación y calidad de datos

Con el objetivo de garantizar la confiabilidad de la información utilizada, se aplican controles de calidad orientados a verificar la consistencia e integridad de los datos. Las principales reglas de validación consideradas son:

- Validación de años dentro del período analizado.
- Verificación de meses válidos.
- Control de valores negativos en montos recaudados.
- Validación de categorías de tipo de contribuyente.
- Consistencia entre provincia y cantón.
- Detección de registros duplicados.
- Identificación de valores nulos en variables clave.
- Verificación de categorías económicas válidas.

Adicionalmente, se implementarán controles específicos de calidad orientados a verificar:

- Porcentaje de registros nulos en variables clave. (% nulos en llave = 0.).
- Porcentaje de registros duplicados respecto a la llave de agregación definida. (% duplicados en llave = 0).
- Presencia de categorías no esperadas dentro de las clasificaciones económicas.
- Consistencia de los valores de recaudación.
- Monitoreo de registros con valores iguales a cero, sin considerarlos necesariamente errores.

Estos controles permitirán construir una base consolidada homogénea y confiable para el desarrollo de los procesos posteriores de análisis exploratorio, segmentación y detección de anomalías.

3.8 Síntesis del capítulo

El capítulo estableció el marco de gobernanza y entendimiento de datos que sustenta el desarrollo metodológico de la investigación. Se identificaron las principales fuentes públicas utilizadas, se describió la función de los diccionarios de datos, se definieron criterios de normalización y se delimitó el alcance analítico de la información disponible.

Asimismo, se precisó que la investigación trabaja con datos agregados y no con registros individuales de contribuyentes. Esta condición define la naturaleza de los resultados: los modelos posteriores podrán identificar patrones y perfiles agregados de interés analítico, pero no eventos individuales de evasión fiscal. En consecuencia, el capítulo proporciona la base conceptual y técnica necesaria para las etapas de preparación de datos, construcción de variables, análisis exploratorio y modelado no supervisado.

Referencia citada en el capítulo

Ministerio de Telecomunicaciones y de la Sociedad de la Información. (2022). *Política de Datos Abiertos del Ecuador. Acuerdo Ministerial MINTEL-MINTEL-2022-0021*.

CAPÍTULO 4

PREPARACIÓN Y LIMPIEZA DE DATOS

Este capítulo describe el proceso de preparación, consolidación y refinamiento de la información utilizada para el desarrollo de la presente investigación. Debido a que el estudio tiene como objetivo identificar patrones asociados al riesgo tributario por medio de técnicas de ciencia de datos, fue necesario recopilar, organizar y armonizar información proveniente de múltiples fuentes abiertas publicadas por el Servicio de Rentas Internas (SRI).

Entre las fuentes consideradas se encuentran los registros de recaudación tributaria, ventas declaradas, catastro de contribuyentes, empresas fantasmas, actividad económica, presión fiscal, comprobantes electrónicos, agregadores de pago, mercados en línea y la dinámica empresarial reflejada a través de las aperturas e inactivaciones de contribuyentes. Cada uno de estos conjuntos de datos aporta una perspectiva diferente y complementaria sobre el comportamiento económico y tributario de los distintos sectores productivos del país, permitiendo construir una visión más integral del fenómeno analizado.

Debido a que la información proviene de fuentes heterogéneas, con diferentes periodos de publicación, niveles de agregación y estructuras de datos, fue necesario aplicar procesos de consolidación, estandarización y validación (Koukaras & Tjortjis, 2025). Estas actividades permitieron construir una base analítica consistente, adecuada para realizar comparaciones temporales, sectoriales y geográficas. Esta etapa constituye un elemento fundamental para garantizar la calidad y confiabilidad de los análisis posteriores, incluyendo los procesos de segmentación, análisis exploratorio y detección de comportamientos atípicos.

Como resultado de este proceso se obtendrá una estructura de datos integrada que servirá como base para la construcción de variables analíticas y la aplicación de modelos de aprendizaje no supervisado. El propósito final consiste en identificar perfiles empresariales y tributarios con características diferenciadas que aporten evidencia estadística útil para el

análisis del riesgo tributario. Los resultados deberán interpretarse como señales, tendencias o segmentos de interés analítico y no como evidencia concluyente de evasión fiscal.

4.1 Fuentes de información utilizada.

En esta investigación trabajamos con distintos conjuntos de datos publicados por el Servicio de Rentas Internas (SRI) dentro la iniciativa de datos abiertos que poseen las entidades Estatales. La selección de estas fuentes fue producto en primera instancia de su validez ya que es proporcionada por la entidad de control misma y por su capacidad para mostrar diferentes dimensiones del comportamiento económico, empresarial con respecto a la tributación.

Entre la información utilizada se incluyen registros de recaudación tributaria, ventas declaradas, catastro de contribuyentes, actividad económica, empresas fantasmas, presión fiscal, comprobantes electrónicos, agregadores de pago, mercados en línea y datos sobre aperturas e inactivaciones empresariales. Cada fuente aporta las piezas del rompecabezas a construir para el análisis, ofreciendo perspectivas directas y/o complementarias que enriquecen la comprensión del fenómeno económico y tributario nacional.

Esta diversidad de datos será clave más adelante, ya que permitirá construir variables analíticas sólidas y facilitar la identificación de patrones relevantes para el estudio del riesgo tributario.

Tabla 5

Fuentes de información utilizadas en la investigación

Fuente de información	Descripción general	Uso dentro de la investigación
Recaudación tributaria	Información histórica de recaudación por impuesto, provincia y actividad económica.	Análisis del comportamiento tributario y comparación sectorial.
Ventas declaradas	Información agregada de ventas reportadas por los contribuyentes.	Medición de actividad económica y comportamiento empresarial.
Catastro RUC	Información de contribuyentes, actividades económicas y ubicación geográfica.	Clasificación y segmentación de perfiles empresariales.
Empresas fantasmas	Registros publicados por el SRI sobre contribuyentes catalogados como empresas fantasmas.	Referencia pública para el análisis de patrones asociados al riesgo tributario.
IAENP	Índice de Actividad Empresarial No Petrolera.	Contextualización del comportamiento económico nacional.

Fuente de información	Descripción general	Uso dentro de la investigación
Presión fiscal	Indicadores agregados de carga tributaria y contribución al Estado.	Análisis del contexto tributario y económico.
Aperturas e inactivaciones	Registros relacionados con la creación y cierre de actividades económicas.	Evaluación de la dinámica empresarial por sector y territorio.
Comprobantes electrónicos	Información asociada a la emisión de comprobantes electrónicos.	Indicadores de formalización tributaria y adopción tecnológica.
Agregadores de pago	Información relacionada con plataformas de procesamiento de pagos.	Caracterización de actividades económicas digitales.
Mercados en línea	Información de plataformas de comercio electrónico registradas.	Complemento para el análisis de la economía digital.

Nota. Elaboración propia a partir de información pública publicada por el Servicio de Rentas Internas (SRI).

4.2 Integración y consolidación de información

4.2.1 Consolidación histórica de la información

Los conjuntos de datos utilizados en la presente investigación fueron publicados por el Servicio de Rentas Internas (SRI) en diferentes períodos de tiempo y son archivos independientes. Por esta razón, el primer paso consistió en realizar un proceso de consolidación histórica, integrando la información disponible para cada fuente dentro de una estructura única que permitiera analizar su evolución a través del tiempo. Este procedimiento facilitó la construcción de series históricas consistentes y comparables para los distintos años considerados en el estudio.

4.2.2 Homologación de estructuras

Una vez consolidada la información histórica de cada fuente, se procedió a realizar un proceso de homologación de estructuras la cual tiene como objetivo de garantizar la consistencia de los datos de los datasets analizados y sean concordantes al periodo de tiempo del cual fueron originarios. Debido a que los conjuntos de información fueron publicados en distintos años, fue necesario verificar la uniformidad de nombres de variables o campos, formato en el que fueron almacenados y tipos de datos utilizados (Roy & Dutta, 2022).

Durante esta etapa se revisaron aspectos relacionados a como se representan las fechas, formatos numéricos, codificaciones territoriales y estructuras de identificadores utilizadas en los distintos conjuntos de datos. De igual manera, se verificó la correspondencia entre variables equivalentes que podían presentar diferencias menores derivadas de cambios en los procesos de publicación o actualización de la información.

Como resultado de este procedimiento se obtuvieron estructuras homogéneas y comparables entre períodos temporales, permitiendo que los análisis a posterior se desarrollen sobre una base sólidas y reduciendo posibles inconsistencias se deriven de diferencias de formato entre los formatos de archivos utilizados.

4.2.3 Entorno tecnológico para preparación de datos

El proceso de extracción, transformación y carga (ETL) de la información fue desarrollado utilizando el lenguaje de programación Python, debido a su amplia adopción en proyectos de análisis de datos y ciencia de datos (Xie & Wan, 2023). La selección de esta tecnología permitió automatizar tareas de consolidación, limpieza y validación sobre grandes volúmenes de información provenientes de distintas fuentes publicadas por el Servicio de Rentas Internas (SRI).

Para la manipulación y transformación de datos se utilizó la biblioteca Pandas, la cual proporciona estructuras eficientes para la lectura, consolidación y procesamiento de información tabular. Adicionalmente, se empleó NumPy para operaciones numéricas y cálculos estadísticos requeridos durante las etapas de validación y construcción de variables (Nelli, 2023).

El desarrollo y ejecución de los procesos se realizó utilizando Visual Studio Code (VS Code) como entorno principal de trabajo. Esta herramienta permitió organizar los scripts de procesamiento, facilitar la depuración de errores y mantener la trazabilidad de las transformaciones aplicadas a cada conjunto de datos.

Las herramientas seleccionadas permitieron implementar un flujo de trabajo reproducible, facilitando la integración de información histórica, la detección de inconsistencias y la preparación de los conjuntos de datos que posteriormente serán utilizados en las etapas de análisis exploratorio y modelado.

4.2.4 Limpieza y tratamiento de inconsistencias

Una vez consolidadas e integradas las diferentes fuentes de información, se inició el proceso de limpieza y validación de datos. Esta etapa tuvo como objetivo identificar posibles inconsistencias, errores de formato, registros incompletos y otros elementos que pudieran afectar la calidad de los análisis posteriores. Debido a que cada conjunto de datos presenta características particulares, la limpieza se realizó de manera independiente para cada fuente, aplicando criterios acordes a su estructura y naturaleza.

4.2.4.1 Proceso de limpieza de la información de recaudación tributaria

Durante el proceso de integración histórica se identificó una diferencia en la codificación de caracteres utilizada en los archivos publicados por el SRI. Los conjuntos de datos comprendidos entre 2017 y 2022 pudieron ser procesados utilizando codificación UTF-8, mientras que los archivos correspondientes al período 2023–2026 requirieron mecanismos alternativos de lectura debido a incompatibilidades de codificación detectadas durante el proceso ETL.

Con el propósito de verificar la consistencia de la información histórica de recaudación tributaria, se realizó una revisión inicial de los archivos correspondientes al período 2017–2026. Esta validación consideró el número de registros, la cantidad de variables disponibles y la codificación utilizada en cada archivo, permitiendo identificar posibles diferencias estructurales que pudieran afectar el proceso de integración de la información.

Tabla 6*Resumen de validación inicial de los archivos de recaudación tributaria (2017–2026)*

Año	Número de registros	Número de variables	Codificación
2017	576.472	11	UTF-8
2018	612.562	11	UTF-8
2019	648.551	11	UTF-8
2020	582.719	11	UTF-8
2021	686.087	11	UTF-8
2022	735.165	11	UTF-8
2023	690.589	11	Latin1
2024	665.184	11	Latin1
2025	696.289	11	Latin1
2026	243.928	11	Latin1

Nota. Elaboración propia a partir de los conjuntos de datos de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). La validación permitió verificar la consistencia estructural de los archivos históricos e identificar diferencias en la codificación de caracteres utilizadas entre los distintos períodos analizados.

Los resultados presentados en la Tabla 6 evidencian que la estructura de los archivos de recaudación tributaria se mantuvo consistente durante todo el período analizado, conservando once variables en cada conjunto de datos. Esta característica facilitó el proceso de integración histórica al no requerir modificaciones significativas en la estructura de los registros.

A pesar de que, durante la etapa de carga de información se identificó una diferencia en la codificación de caracteres utilizada en los archivos publicados por el SRI. Los conjuntos correspondientes al período 2017–2022 pudieron ser procesados utilizando codificación UTF-8, mientras que los archivos publicados entre 2023 y 2026 requirieron el uso de codificación Latin1 para garantizar su correcta lectura. Esta situación fue incorporada dentro del proceso ETL mediante mecanismos de carga adaptativos que permitieron homogenizar el tratamiento de la información sin afectar la integridad de los datos.

Una vez superadas estas diferencias técnicas, se procedió a verificar la consistencia de las variables disponibles en cada período con el objetivo de asegurar la comparabilidad de la información antes de iniciar las transformaciones y validaciones específicas de los datos de recaudación tributaria.

Adicionalmente, se verificó que los archivos correspondientes al período 2017–2026 conservaron la misma estructura de variables y el mismo orden de campos, permitiendo realizar la integración histórica sin requerir transformaciones estructurales adicionales. Esta consistencia facilitó la construcción de una base consolidada de recaudación tributaria para su posterior análisis.

Tabla 7

Validación de consistencia estructural y temporal de la información de recaudación tributaria (2017-2026)

Año	Número de registros	Número de variables	Codificación	Estructura de variables	Cobertura temporal
2017	576.472	11	UTF-8	Consistente	Completa (12 meses)
2018	612.562	11	UTF-8	Consistente	Completa (12 meses)
2019	648.551	11	UTF-8	Consistente	Completa (12 meses)
2020	582.719	11	UTF-8	Consistente	Completa (12 meses)
2021	686.087	11	UTF-8	Consistente	Completa (12 meses)
2022	735.165	11	UTF-8	Consistente	Completa (12 meses)
2023	690.589	11	Latin1	Consistente	Completa (12 meses)
2024	665.184	11	Latin1	Consistente	Completa (12 meses)
2025	696.289	11	Latin1	Consistente	Completa (12 meses)
2026	243.928	11	Latin1	Consistente	Parcial (4 meses)

La validación permitió verificar la consistencia estructural de los archivos históricos, identificar diferencias en la codificación de caracteres y determinar la cobertura temporal disponible para cada período analizado.

Los resultados obtenidos evidencian que los archivos de recaudación tributaria mantuvieron una estructura homogénea durante todo el período analizado, conservando once variables con la misma definición y orden de campos. La principal diferencia identificada corresponde a la codificación de caracteres utilizada en los archivos publicados a partir de 2023, situación que fue corregida durante el proceso ETL mediante mecanismos de lectura adaptativos. Asimismo, se verificó que la información correspondiente a los años 2017–2025

presenta cobertura completa de doce meses, mientras que el año 2026 dispone únicamente de información parcial para cuatro meses, condición que deberá considerarse en los análisis temporales posteriores.

4.2.4.2 *Evaluación de calidad de la información de recaudación tributaria*

La calidad de los datos constituye un elemento fundamental en cualquier proceso de análisis estadístico y ciencia de datos. La presencia de valores nulos, registros duplicados o inconsistencias en la información puede introducir sesgos en los resultados, afectar la representatividad de los datos y generar interpretaciones erróneas sobre los fenómenos estudiados. Debido a ello, una vez consolidada la información histórica de recaudación tributaria, se realizó una evaluación de calidad orientada a verificar la integridad y consistencia de los registros, garantizando que la información utilizada en las etapas posteriores de análisis y modelado refleje adecuadamente las características reales del conjunto de datos.

Tabla 8
Resultados de la validación inicial de calidad de los registros de recaudación tributaria

Criterio evaluado	Resultado obtenido
Registros analizados	6.137.546
Variables evaluadas	11
Valores nulos	0
Variables con datos faltantes	0
Cobertura temporal completa	2017–2025
Cobertura temporal parcial	2026 (4 meses)
Consistencia de variables	Validada
Consistencia de categorías	Validada

Nota. Elaboración propia a partir del conjunto consolidado de recaudación tributaria obtenido durante el proceso ETL.

Los resultados obtenidos evidencian un elevado nivel de calidad en la información analizada. La revisión realizada no identificó valores nulos en ninguna de las variables evaluadas, por lo que no fue necesario aplicar técnicas de imputación, eliminación de registros o sustitución de datos. De igual manera, la validación de registros duplicados no evidenció observaciones repetidas dentro del conjunto consolidado, garantizando que cada registro aporte información única al análisis. Asimismo, las categorías evaluadas presentaron consistencia en

su representación a lo largo del período estudiado, permitiendo preservar la totalidad de la información disponible para las etapas posteriores de exploración, construcción de variables y modelado.

Además, la evaluación de calidad realizada sobre el conjunto consolidado de recaudación tributaria evidenció un alto nivel de integridad de la información. El conjunto final quedó conformado por 6.137.546 registros y 11 variables, sin presentar valores nulos ni registros duplicados. Asimismo, las variables categóricas analizadas mantuvieron consistencia en su representación a lo largo de todo el período estudiado.

La validación territorial permitió identificar veinticinco categorías dentro de la variable provincia, de las cuales veinticuatro corresponden a divisiones geográficas oficiales del Ecuador y una categoría adicional denominada “NO TIENE”, utilizada para representar registros sin asignación territorial específica. Por otra parte, se verificó la existencia de información para los doce meses de cada año entre 2017 y 2025, mientras que el año 2026 dispone únicamente de información correspondiente a cuatro meses, situación que deberá considerarse en los análisis temporales posteriores.

Los resultados obtenidos permiten concluir que la información de recaudación tributaria presenta condiciones adecuadas de calidad para su utilización en las etapas posteriores de análisis exploratorio, construcción de variables analíticas y modelado.

4.2.4.3 Proceso de limpieza de la información de compras y ventas declaradas

Durante la inspección inicial de los archivos históricos de compras y ventas declaradas se identificaron diferencias relevantes en la estructura y formato de publicación de la información. Los archivos correspondientes al período 2017–2024 presentan una estructura homogénea de catorce variables, utilizando barra vertical como separador de campos y codificación Latin1. En cambio, el archivo correspondiente al año 2025 utiliza coma como

separador y presenta una estructura ampliada de dieciséis variables. El archivo 2026 mantiene la estructura ampliada, aunque utiliza nuevamente barra vertical como separador.

La diferencia estructural observada a partir de 2025 se relaciona principalmente con una mayor desagregación de las ventas y compras gravadas, incorporando campos como ventas con tarifa gravada, tarifa variable y tarifa 5 %, además de compras netas con tarifa gravada. Debido a este cambio, la consolidación histórica de esta fuente requiere un proceso de homologación de campos antes de construir indicadores comparables.

Con el objetivo de mantener la consistencia temporal del análisis, se priorizó el uso de variables comunes presentes en todos los archivos revisados: año, mes, código de sector económico, provincia, cantón, exportaciones, importaciones, total de compras y total de ventas. Estas variables permiten construir indicadores generales de actividad económica sin que los cambios en la desagregación por tarifa afecten la comparabilidad de la serie histórica.

Tabla 9
Resumen de validación estructural de los archivos de compras y ventas declaradas, período 2017–2026

Año	Registros	Variables	Separador	Codificación	Estructura	Cobertura
2017	50.223	14		Latin1	Antigua	Completa
2018	50.334	14		Latin1	Antigua	Completa
2019	50.366	14		Latin1	Antigua	Completa
2020	49.903	14		Latin1	Antigua	Completa
2021	49.908	14		Latin1	Antigua	Completa
2022	48.977	14		Latin1	Antigua	Completa
2023	47.996	14		Latin1	Antigua	Completa
2024	49.852	14		Latin1	Antigua	Completa
2025	47.784	16	,	UTF-8-SIG	Ampliada	Completa
2026	8.439	16		Latin1	Ampliada	Parcial

Nota. Elaboración propia a partir de los archivos históricos de compras y ventas declaradas publicados por el Servicio de Rentas Internas (SRI).

La estructura antigua corresponde a los archivos con catorce variables publicados entre 2017 y 2024, mientras que la estructura ampliada corresponde a los archivos con dieciséis variables observados desde 2025. El año 2026 se considera parcial debido a que no dispone de cobertura anual completa al momento de la investigación.

Los resultados de la Tabla 9 muestran que los archivos de compras y ventas declaradas mantienen una estructura homogénea durante el período 2017–2024, con catorce variables,

separador de barra vertical y codificación Latin1. Esta consistencia facilita la integración histórica de los primeros años de la serie.

A partir de 2025 se identifica un cambio estructural en la fuente, pasando de catorce a dieciséis variables. Este cambio responde a una mayor desagregación de los campos relacionados con ventas y compras gravadas. Además, el archivo correspondiente a 2025 utiliza coma como separador, a diferencia del resto de años revisados, lo que requirió un tratamiento específico durante la etapa de lectura de datos.

El archivo 2026 conserva la estructura ampliada, pero presenta cobertura parcial, por lo que no será utilizado en comparaciones anuales completas. En función de estos hallazgos, la consolidación histórica de esta fuente se realizará mediante una homologación de campos comunes, priorizando variables comparables como año, mes, sector económico, provincia, cantón, exportaciones, importaciones, total de compras y total de ventas.

Tabla 10

Resultados de la homologación y validación de calidad de compras y ventas declaradas, período 2017–2026

Año	Registros	Variables homologadas	Nulos totales	Nulos en código de sector	Duplicados	Meses disponibles	Separador	Codificación
2017	50.223	9	0	0	0	12		Latin1
2018	50.334	9	0	0	0	12		Latin1
2019	50.366	9	0	0	0	12		Latin1
2020	49.903	9	0	0	0	12		Latin1
2021	49.908	9	0	0	0	12		Latin1
2022	48.977	9	0	0	0	12		Latin1
2023	47.996	9	0	0	0	12		Latin1
2024	49.852	9	0	0	0	12		Latin1
2025	47.784	9	96	96	0	12	,	UTF-8-SIG
2026	8.439	9	0	0	0	2		Latin1

Nota. Elaboración propia a partir de los archivos históricos de compras y ventas declaradas publicados por el

Servicio de Rentas Internas (SRI).

La tabla resume la validación de calidad realizada sobre la base homologada, construida a partir de variables comunes presentes en todos los años analizados. Los valores nulos identificados corresponden exclusivamente a la variable código de sector económico en el año

2025. El año 2026 se considera parcial debido a que dispone únicamente de dos meses de información.

Los resultados de la Tabla 10 muestran que, luego del proceso de homologación, la base de compras y ventas declaradas quedó estructurada con nueve variables comunes para todos los años analizados. Esta selección permitió integrar información proveniente de distintas estructuras de publicación sin perder comparabilidad temporal, priorizando campos relevantes para el análisis posterior, como año, mes, sector económico, provincia, cantón, exportaciones, importaciones, total de compras y total de ventas.

La validación de calidad evidencia que los archivos correspondientes al período 2017–2024 no presentan valores nulos ni registros duplicados en las variables homologadas. En el año 2025 se identificaron 96 valores nulos, todos asociados a la variable CODIGO_SECTOR_N1, mientras que el resto de las variables comunes no presentó datos faltantes. Por su parte, el año 2026 no registra valores nulos ni duplicados, aunque dispone únicamente de dos meses de información, por lo que se mantiene como período parcial.

Debido a que los 96 registros con código de sector económico faltante presentaban información válida en las variables de compras, ventas, provincia y cantón, no fueron eliminados de la base consolidada. En su lugar, la variable CODIGO_SECTOR_N1 fue homologada mediante la categoría “NO CLASIFICADO”, con el fin de conservar la información económica disponible y evitar la pérdida de registros útiles para los análisis nacionales, temporales y territoriales. Luego de este tratamiento, la base homologada no presentó valores nulos en ninguna de las variables comunes seleccionadas.

En conjunto, los resultados permiten concluir que la base de compras y ventas declaradas presenta condiciones adecuadas para continuar con la construcción de variables analíticas relacionadas con ventas, compras, exportaciones, importaciones y balance económico declarado. No obstante, la categoría “NO CLASIFICADO” deberá ser considerada

de forma diferenciada en los análisis sectoriales posteriores, mientras que el año 2026 deberá tratarse con cautela debido a su cobertura parcial.

CAPÍTULO 5

CONSTRUCCIÓN DE VARIABLES ANALÍTICAS

El presente capítulo describe el proceso de construcción de variables analíticas a partir de los conjuntos de datos previamente consolidados y depurados. Una vez completadas las etapas de integración, validación y limpieza de la información, fue posible transformar los datos originales en indicadores que permitan representar de manera más adecuada los fenómenos económicos y tributarios objeto de estudio.

Las variables construidas tienen como finalidad resumir características relevantes del comportamiento de los contribuyentes, la dinámica económica de los distintos sectores productivos y la evolución temporal de los indicadores tributarios. Para ello, se generaron métricas derivadas relacionadas con niveles de actividad, participación relativa, crecimiento, concentración y variabilidad, utilizando información proveniente de las diferentes fuentes analizadas.

La construcción de estas variables constituye una etapa fundamental dentro del proceso de ciencia de datos, ya que permite transformar grandes volúmenes de información en atributos con capacidad descriptiva y analítica. Los indicadores obtenidos servirán como insumo para las etapas posteriores de análisis exploratorio, identificación de patrones y aplicación de técnicas de aprendizaje no supervisado orientadas al estudio del riesgo tributario.

5.1 Metodología para la construcción de variables analíticas

Una vez concluido el proceso de integración, depuración y validación de la información descrito en el Capítulo 4, se procedió a la construcción de variables analíticas orientadas a representar de manera más adecuada los fenómenos económicos y tributarios considerados en la investigación.

El proceso de construcción de variables tuvo como objetivo transformar los datos originales en indicadores capaces de resumir comportamientos relevantes, facilitar la comparación entre distintos períodos de tiempo y proporcionar atributos útiles para las etapas posteriores de análisis exploratorio y modelado. Para ello, se aplicaron técnicas de agregación, cálculo de participaciones relativas, análisis de crecimiento, medición de concentración y estimación de indicadores de dispersión (Forke & Tropmann-Frick, 2021).

Las variables generadas fueron construidas a partir de las diferentes fuentes de información utilizadas en el estudio, incluyendo recaudación tributaria, ventas declaradas, actividad económica, catastro de contribuyentes y demás conjuntos de datos incorporados en la investigación. Dependiendo de la naturaleza de cada fuente, se desarrollaron indicadores orientados a describir la evolución temporal de los fenómenos observados, su distribución geográfica, la participación de determinados sectores económicos y la identificación de patrones de concentración o variabilidad.

De manera general, las variables construidas pueden clasificarse en cinco categorías principales: variables de agregación, variables de participación relativa, variables de crecimiento temporal, variables de concentración e indicadores de dispersión. Esta clasificación permitió organizar de forma sistemática los diferentes indicadores obtenidos y facilitar su posterior utilización en las etapas de análisis exploratorio y detección de patrones asociados al riesgo tributario.

Las siguientes secciones describen las variables construidas para cada una de las fuentes de información analizadas, detallando su procedimiento de cálculo y la utilidad analítica que aportan dentro del estudio.

5.2 Variables derivadas de la recaudación tributaria

La información de recaudación tributaria constituye una de las principales fuentes analizadas dentro de la investigación, debido a su capacidad para reflejar el comportamiento de los ingresos fiscales a nivel temporal, geográfico y tributario.

A partir del conjunto consolidado correspondiente al período 2017–2026, se construyeron diversos indicadores orientados a caracterizar la evolución de la recaudación nacional, su distribución territorial y la participación de los diferentes componentes tributarios dentro del total recaudado.

Las variables presentadas en esta sección permiten describir tendencias generales de la recaudación y establecer una base analítica para las etapas posteriores de análisis exploratorio e identificación de patrones asociados al riesgo tributario.

5.2.1 Recaudación tributaria total anual

La recaudación tributaria total anual constituye uno de los principales indicadores para analizar la evolución de los ingresos fiscales a lo largo del tiempo. Esta variable fue obtenida mediante la agregación del valor recaudado para cada año del período analizado, permitiendo identificar tendencias generales, variaciones temporales y cambios en el comportamiento de la recaudación tributaria nacional.

Tabla 11
Recaudación tributaria total anual del Ecuador, período 2017–2026

Año	Recaudación total anual (USD)
2017	13.679.250.000
2018	15.143.840.000
2019	14.268.760.000
2020	12.381.840.000
2021	13.976.160.000
2022	17.161.840.000
2023	17.419.730.000
2024	20.131.290.000
2025	21.500.870.000
2026	8.742.128.000

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). Los valores corresponden a la suma anual de la variable VALOR_RECAUDADO. La información correspondiente al año 2026 es parcial y comprende únicamente los primeros cuatro meses disponibles al momento de la investigación.

Los resultados presentados en la Tabla 11 muestran una tendencia general de crecimiento de la recaudación tributaria nacional durante el período analizado. Entre 2017 y 2018 se observa un incremento de los ingresos fiscales, mientras que en los años 2019 y 2020 se evidencia una reducción, siendo este último el período con la disminución más significativa, situación que puede asociarse a los efectos económicos derivados de la pandemia de COVID-19.

A partir de 2021 se aprecia una recuperación sostenida de la recaudación tributaria, alcanzando sus valores más altos en los años 2024 y 2025. Este comportamiento refleja una mejora progresiva en la actividad económica y en la capacidad de recaudación fiscal observada durante los años posteriores a la emergencia sanitaria. Finalmente, el valor registrado para 2026 debe interpretarse con cautela debido a que corresponde únicamente a información parcial disponible para los primeros meses del año.

5.2.2 Variación interanual de la recaudación tributaria

La variación interanual de la recaudación tributaria permite medir el cambio porcentual de los ingresos fiscales respecto al año anterior. Este indicador facilita la identificación de períodos de crecimiento, contracción o recuperación dentro de la serie histórica analizada.

Tabla 12
Variación interanual de la recaudación tributaria, período 2017–2025

Año	Variación interanual (%)
2017	N/D
2018	10,71
2019	-5,78
2020	-13,22
2021	12,88
2022	22,79
2023	1,50
2024	15,57
2025	6,80

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria del Servicio de Rentas Internas (SRI). La variación interanual corresponde al cambio porcentual de la recaudación total anual respecto al año inmediatamente anterior. Para el año 2017 no fue posible calcular la variación debido a la ausencia de información del período precedente. El año 2026 fue excluido del análisis por corresponder a información parcial disponible únicamente para los primeros meses del año.

Los resultados presentados en la Tabla 12 evidencian que la recaudación tributaria nacional experimentó períodos alternados de crecimiento y contracción a lo largo del horizonte analizado. Entre 2018 y 2019 se observa una desaceleración de la recaudación, situación que se profundiza en 2020 con una reducción del 13,22 %, constituyendo la mayor caída registrada en la serie histórica. Este comportamiento coincide con las afectaciones económicas derivadas de la pandemia de COVID-19 (wgaleas, 2021).

A partir de 2021 se evidencia una recuperación sostenida de los ingresos tributarios, destacándose el crecimiento de 22,79 % registrado en 2022, el valor más alto de la serie analizada. Durante los años 2023, 2024 y 2025 la recaudación mantiene una tendencia positiva, aunque con ritmos de crecimiento más moderados. En conjunto, los resultados reflejan una capacidad de recuperación significativa de la recaudación tributaria nacional después de la contracción observada en 2020.

5.2.3 Recaudación tributaria total mensual

La recaudación tributaria total mensual permite analizar el comportamiento de los ingresos fiscales con un mayor nivel de detalle temporal. Esta variable fue obtenida mediante la agregación del valor recaudado para cada combinación de año y mes del período analizado, permitiendo identificar patrones recurrentes, variaciones estacionales y cambios en la dinámica de la recaudación tributaria a lo largo del tiempo.

Tabla 13*Participación mensual de la recaudación tributaria por año, período 2017–2025*

Mes	2017	2018	2019	2020	2021	2022	2023	2024	2025	Total
Enero	10,09	9,02	9,48	11,46	9,33	9,07	9,67	8,09	9,38	9,41
Febrero	6,29	6,31	6,90	7,34	6,76	6,04	6,45	5,88	6,37	6,43
Marzo	8,53	7,38	7,63	9,35	8,59	10,92	9,82	8,54	9,30	8,95
Abril	12,90	11,75	14,49	10,87	10,19	12,76	13,77	12,18	10,39	12,12
Mayo	7,76	8,09	7,88	5,92	7,61	7,75	7,81	8,74	7,62	7,76
Junio	7,91	6,80	7,27	6,17	7,54	7,35	7,64	7,64	7,44	7,34
Julio	8,21	7,96	8,04	7,10	7,87	7,23	7,35	10,22	8,04	8,07
Agosto	7,33	7,19	8,00	7,30	7,68	7,52	7,46	7,81	8,04	7,62
Septiembre	8,18	7,76	7,72	10,20	8,77	8,14	7,54	7,81	8,23	8,19
Octubre	7,38	7,26	7,78	7,72	8,32	7,61	7,36	7,48	7,86	7,64
Noviembre	7,61	7,28	7,57	7,97	8,40	7,58	7,09	7,68	8,33	7,73
Diciembre	7,81	13,22	7,25	8,62	8,94	8,03	8,04	7,92	9,01	8,74

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el

Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual de cada mes respecto al total anual de recaudación. La columna Total corresponde a la participación de cada mes sobre la recaudación acumulada del período 2017–2025. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

Los resultados presentados en la Tabla 13 permiten observar que la recaudación tributaria no se distribuye de manera uniforme a lo largo del año. Abril concentra la mayor participación acumulada del período, con 12,12 %, seguido por enero, marzo y diciembre, meses que también muestran un peso relativo importante dentro de la recaudación anual. Este comportamiento evidencia la existencia de una dinámica mensual diferenciada, donde ciertos períodos concentran una proporción mayor de los ingresos tributarios.

Febrero presenta la menor participación acumulada, con 6,43 %, mientras que la mayoría de los meses restantes se mantienen en rangos relativamente cercanos, entre aproximadamente 7 % y 8 %. Esta distribución permite identificar un patrón mensual útil para el análisis posterior, ya que las desviaciones respecto a estos comportamientos esperados podrían ser relevantes al momento de estudiar variaciones atípicas o cambios en la dinámica de recaudación.

También se observan comportamientos particulares en algunos años, como el mayor peso relativo de enero, marzo y septiembre en 2020, así como la participación elevada de diciembre en 2018. Estos comportamientos sugieren que el análisis mensual aporta

información adicional frente a una revisión exclusivamente anual, permitiendo captar variaciones temporales que podrían pasar desapercibidas en indicadores más agregados.

5.2.4 Participación provincial en la recaudación tributaria

La participación provincial en la recaudación tributaria permite identificar el peso relativo que tiene cada territorio dentro del total recaudado a nivel nacional. Esta variable resulta relevante porque la actividad económica y tributaria no se distribuye de manera uniforme entre provincias, sino que tiende a concentrarse en determinados territorios con mayor dinamismo productivo, comercial y empresarial.

Para construir este indicador se agregó el valor recaudado por provincia y año, y posteriormente se calculó la participación porcentual de cada provincia respecto al total nacional del mismo período. Este procedimiento permite comparar la importancia relativa de cada territorio dentro de la recaudación anual y observar posibles cambios en la concentración geográfica de los ingresos tributarios.

Tabla 14

Participación provincial en la recaudación tributaria, principales provincias, período 2017–2025

Provincia	2017	2018	2019	2020	2021	2022	2023	2024	2025	Total
Pichincha	53,10	54,88	52,25	50,64	49,58	49,45	50,17	50,63	49,23	50,96
Guayas	30,30	28,98	31,06	33,45	32,49	31,15	30,22	30,94	31,11	31,02
Azuay	4,97	4,98	4,89	4,59	5,23	5,18	4,98	4,66	4,52	4,87
Zamora Chinchipe	0,55	0,68	0,77	0,52	0,98	2,41	2,98	3,06	3,99	1,99
Manabí	1,91	1,88	1,84	1,87	2,01	1,98	1,97	1,98	2,10	1,96
Tungurahua	1,66	1,59	1,78	1,56	1,72	1,74	1,63	1,45	1,62	1,63
El Oro	1,43	1,43	1,45	1,49	1,72	1,78	1,70	1,44	1,57	1,56
Cotopaxi	0,85	0,85	0,84	0,85	0,98	0,96	0,90	0,93	0,93	0,91
Imbabura	0,75	0,72	0,78	0,79	0,79	0,78	0,73	0,61	0,60	0,72
Los Ríos	0,56	0,52	0,57	0,66	0,63	0,60	0,58	0,58	0,62	0,59

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el

Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual de cada provincia

respecto al total anual de recaudación. La columna Total corresponde a la participación acumulada de cada

provincia sobre el total recaudado durante el período 2017–2025. Se presentan las diez provincias con mayor participación acumulada.

Los resultados de la Tabla 14 evidencian una alta concentración territorial de la recaudación tributaria. Pichincha y Guayas concentran conjuntamente más del 80 % de la recaudación acumulada durante el período 2017–2025, lo que refleja el peso fiscal de los principales polos económicos y administrativos del país. Azuay se ubica como la tercera provincia con mayor participación, aunque con una diferencia considerable frente a las dos primeras.

Un comportamiento relevante se observa en Zamora Chinchipe, cuya participación aumenta de manera sostenida a partir de 2021, pasando de valores inferiores al 1 % en los primeros años del período a 3,99 % en 2025. Este cambio sugiere una modificación importante en la distribución territorial de la recaudación, que podría estar vinculada con dinámicas sectoriales específicas y que deberá ser revisada con mayor detalle en el análisis exploratorio posterior.

En conjunto, esta variable permite identificar patrones de concentración geográfica y constituye un insumo relevante para la construcción de perfiles analíticos territoriales. Su incorporación facilita comparar provincias con alta concentración tributaria frente a territorios con menor participación relativa, así como observar variaciones inusuales en la distribución espacial de la recaudación.

5.2.5 Participación por grupo de impuesto

La participación por grupo de impuesto permite identificar qué componentes tributarios concentran la mayor proporción de la recaudación nacional. Esta variable resulta relevante porque no todos los impuestos tienen el mismo peso dentro de la estructura fiscal, por lo que analizar su distribución permite comprender qué grupos tributarios sostienen en mayor medida los ingresos recaudados durante el período de estudio.

Para construir este indicador se agrupó el valor recaudado por año y grupo de impuesto. Posteriormente, se calculó la participación porcentual de cada grupo respecto al total anual de

recaudación. Este procedimiento permite observar la composición tributaria del país y detectar cambios en el peso relativo de los distintos grupos de impuestos a lo largo del tiempo.

Tabla 15

Participación por grupo de impuesto en la recaudación tributaria, período 2017–2025

Grupo de impuesto	2017	2018	2019	2020	2021	2022	2023	2024	2025	Total
Impuesto al Valor Agregado	49,19	44,50	46,85	44,47	48,13	47,39	48,48	49,90	51,52	48,12
Impuesto a la Renta Global	30,54	35,13	33,43	35,59	30,99	30,59	33,37	32,98	31,47	32,59
Salida de Divisas	8,02	7,96	7,99	7,79	8,67	7,42	6,25	6,32	5,97	7,23
Impuesto a los Consumos Especiales	6,94	6,51	6,38	5,98	5,94	4,99	4,75	3,75	3,46	5,22
Contribuciones	1,02	0,76	0,89	2,62	2,68	5,44	3,33	3,76	2,74	2,71
Impuesto a los Vehículos Motorizados	1,40	1,42	1,56	1,56	1,56	1,34	1,48	1,30	1,23	1,41
Impuestos Mineros	0,39	0,37	0,47	0,33	0,53	1,18	1,19	1,14	1,44	0,85
Impuestos de Fomento Ambiental	1,04	1,02	1,11	0,29	0,30	0,29	0,28	0,23	0,20	0,49

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el

Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual de cada grupo de impuesto respecto al total anual de recaudación. La columna Total corresponde a la participación acumulada de cada grupo sobre el total recaudado durante el período 2017–2025. Se presentan los ocho grupos de impuesto con mayor participación acumulada.

Los resultados de la Tabla 15 evidencian que la estructura de la recaudación tributaria se concentra principalmente en dos grupos de impuestos: el Impuesto al Valor Agregado y el Impuesto a la Renta Global. En conjunto, ambos representan más del 80 % de la recaudación acumulada durante el período 2017–2025, lo que confirma su papel central dentro de la estructura fiscal nacional.

El Impuesto al Valor Agregado mantiene la mayor participación durante todo el período analizado, con una participación acumulada de 48,12 %. Por su parte, el Impuesto a la Renta Global registra una participación acumulada de 32,59 %, ubicándose como el segundo componente de mayor peso. Esta composición muestra que la recaudación nacional depende principalmente de impuestos asociados al consumo y a la renta.

También se observa una reducción progresiva en la participación de la Salida de Divisas y del Impuesto a los Consumos Especiales durante los últimos años del período. En contraste, los impuestos mineros muestran un aumento relativo desde 2022, aunque su participación acumulada sigue siendo reducida frente a los principales grupos tributarios. Estos cambios en la composición tributaria pueden aportar señales relevantes para el análisis posterior de patrones sectoriales, territoriales y temporales.

5.2.6 Participación por tipo de contribuyente

La participación por tipo de contribuyente permite analizar cómo se distribuye la recaudación tributaria entre los distintos grupos registrados en la información disponible. Este indicador resulta relevante porque permite diferenciar el peso relativo de sociedades, personas naturales y registros sin clasificación específica dentro del total recaudado.

Para construir esta variable se agrupó el valor recaudado por año y tipo de contribuyente. Posteriormente, se calculó la participación porcentual de cada categoría respecto al total anual de recaudación. Este procedimiento permite observar la estructura de contribución tributaria según el tipo de contribuyente y analizar si dicha composición se mantiene estable o presenta cambios a lo largo del período estudiado.

Tabla 16

Participación por tipo de contribuyente en la recaudación tributaria, período 2017–2025

Tipo de contribuyente	2017	2018	2019	2020	2021	2022	2023	2024	2025	Total
Sociedades	91,21	91,07	90,02	89,53	89,80	90,12	91,28	92,14	92,63	91,02
Personas naturales	7,05	6,66	7,17	6,08	6,74	6,83	5,95	5,45	5,16	6,25
No tiene	1,74	2,28	2,81	4,39	3,46	3,05	2,77	2,41	2,21	2,73

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el

Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual de cada tipo de

contribuyente respecto al total anual de recaudación. La columna Total corresponde a la participación

acumulada de cada categoría sobre el total recaudado durante el período 2017–2025. Se excluye el año 2026

debido a que la información disponible corresponde únicamente a un período parcial.

Los resultados de la Tabla 16 evidencian que la recaudación tributaria se concentra principalmente en las sociedades, las cuales representan el 91,02 % del total acumulado durante el período 2017–2025. Esta participación se mantiene estable a lo largo de toda la serie analizada, con valores superiores al 89 % en todos los años, lo que confirma el peso predominante de este tipo de contribuyente dentro de la estructura tributaria nacional.

Este resultado respalda la decisión metodológica de priorizar el análisis de sociedades y contribuyentes registrados mediante RUC dentro de la investigación, debido a que concentran la mayor parte de la recaudación observada. Asimismo, la marcada diferencia entre sociedades y personas naturales permite justificar la construcción de perfiles analíticos diferenciados según el tipo de contribuyente.

5.2.7 Participación por condición de gran contribuyente

La participación por condición de gran contribuyente permite analizar el peso relativo que tienen los contribuyentes de mayor relevancia fiscal dentro de la recaudación tributaria nacional. Esta variable resulta importante porque los grandes contribuyentes suelen concentrar una proporción significativa de los ingresos fiscales, por lo que su análisis permite observar niveles de concentración tributaria y diferencias frente al resto de contribuyentes.

Para construir este indicador se agrupó el valor recaudado por año y por la variable de gran contribuyente. Posteriormente, se calculó la participación porcentual de cada categoría respecto al total anual de recaudación. Este procedimiento permite identificar la proporción de ingresos tributarios asociada a contribuyentes clasificados como grandes contribuyentes y evaluar su comportamiento a lo largo del período analizado.

Tabla 17

Participación por condición normalizada de gran contribuyente, período 2021–2025

Condición de contribuyente	2021	2022	2023	2024	2025	Total
No gran contribuyente	65,59	51,22	50,60	47,90	48,09	51,84
Gran contribuyente	34,41	48,78	49,40	52,10	51,91	48,16

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual de cada condición de contribuyente respecto al total anual de recaudación. La variable GRAN_CONTRIBUYENTE fue normalizada en dos categorías: “Gran contribuyente” y “No gran contribuyente”. El análisis se limita al período 2021–2025 debido a que en los años anteriores la variable no presenta una desagregación comparable.

La categoría de no grandes contribuyentes mantiene una participación acumulada ligeramente superior durante el período analizado, con 51,84 %, aunque su peso relativo disminuye después de 2021. En conjunto, estos resultados muestran que la condición de gran contribuyente constituye una variable relevante para la segmentación tributaria, ya que permite diferenciar perfiles con distinto nivel de incidencia fiscal dentro de la recaudación nacional.

La normalización aplicada a esta variable permitió evitar interpretaciones erróneas derivadas de cambios en la codificación original de los datos. Por tanto, este indicador será considerado como una dimensión importante en la construcción de perfiles analíticos y en las etapas posteriores de análisis exploratorio y modelado.

5.3 Variables derivadas de compras y ventas declaradas

La información de compras y ventas declaradas permite aproximarse al comportamiento económico reportado por los contribuyentes a través de variables relacionadas con ventas, compras, exportaciones e importaciones. A diferencia de la recaudación tributaria, que refleja los valores efectivamente recaudados por concepto de impuestos, esta fuente permite observar la dinámica económica declarada por los contribuyentes en términos de actividad comercial agregada.

A partir de la base homologada construida en el proceso de preparación de datos, se generaron indicadores orientados a medir el volumen de actividad económica declarada, la relación entre compras y ventas, y el balance económico agregado por período, territorio y sector económico. Estas variables resultan relevantes para la investigación debido a que

permiten analizar posibles diferencias entre ingresos y egresos declarados, así como identificar patrones agregados que podrían aportar señales de interés para el análisis de riesgo tributario.

Las variables derivadas de compras y ventas declaradas complementan la información de recaudación tributaria y permiten observar el fenómeno desde una perspectiva económica más directa. Por esta razón, en esta sección se construyen indicadores anuales y relativos que servirán como insumo para el análisis exploratorio posterior y para la construcción de perfiles analíticos orientados a la detección de comportamientos atípicos.

5.3.1 Compras y ventas declaradas totales por año

Las compras y ventas declaradas totales por año constituyen indicadores base para analizar la evolución de la actividad económica reportada en la información pública del SRI. Estas variables permiten observar el comportamiento agregado de los valores declarados por los contribuyentes, así como identificar períodos de crecimiento, contracción o recuperación dentro de la serie histórica.

Para construir estos indicadores se agruparon los registros de la base homologada por año y se calcularon los totales de ventas, compras, exportaciones e importaciones. Adicionalmente, se construyó el balance declarado, definido como la diferencia entre el total de ventas y el total de compras, y el ratio compras/ventas, calculado como la relación entre las compras y las ventas declaradas de cada período.

Tabla 18

Compras y ventas declaradas totales por año, período 2017–2026

Año	Total ventas	Total compras	Exportaciones	Importaciones	Balance declarado	Ratio compras/ventas
2017	178,27	146,47	19,21	21,77	31,80	0,82
2018	188,85	156,27	20,58	25,43	32,58	0,83
2019	192,20	156,25	21,86	22,03	35,95	0,81
2020	163,32	129,25	19,89	16,44	34,06	0,79
2021	200,03	160,09	26,87	22,76	39,94	0,80
2022	229,26	182,26	31,43	27,75	47,00	0,79
2023	238,37	189,55	30,75	25,72	48,82	0,80
2024	242,99	192,98	34,40	24,53	50,00	0,79
2025	265,98	217,31	37,48	27,89	48,67	0,82
2026	40,68	33,59	6,31	4,54	7,09	0,83

Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). Los valores monetarios se expresan en miles de millones de dólares estadounidenses. El balance declarado corresponde a la diferencia entre total de ventas y total de compras. El ratio compras/ventas se calculó dividiendo el total de compras entre el total de ventas. La información correspondiente al año 2026 es parcial y comprende únicamente los meses disponibles al momento de la investigación.

Los resultados de la Tabla 18 evidencian una tendencia general de crecimiento en las ventas y compras declaradas durante el período 2017–2025, con una caída importante en 2020. Las ventas pasan de 178,27 mil millones de dólares en 2017 a 265,98 mil millones en 2025, mientras que las compras aumentan de 146,47 a 217,31 mil millones en el mismo período. Este comportamiento muestra una expansión de la actividad económica declarada, interrumpida temporalmente durante 2020.

El balance declarado se mantiene positivo en todos los años completos analizados, lo que indica que las ventas totales superan a las compras totales durante toda la serie. Este indicador alcanza sus valores más altos entre 2022 y 2025, con montos cercanos o superiores a 47 mil millones de dólares, reflejando una recuperación posterior a la contracción observada en 2020.

El ratio compras/ventas presenta una relativa estabilidad, con valores entre 0,79 y 0,83. Esta consistencia sugiere que, a nivel agregado, la relación entre compras y ventas declaradas se mantiene dentro de un rango similar durante el período analizado. El año 2026 debe interpretarse con cautela, debido a que corresponde a información parcial y no es directamente comparable con los años completos.

5.3.2 Variación interanual de compras y ventas declaradas

La variación interanual de compras y ventas declaradas permite medir los cambios porcentuales registrados en la actividad económica reportada de un año a otro. Este indicador

facilita la identificación de períodos de expansión, contracción o recuperación en los valores declarados por los contribuyentes, complementando el análisis de los totales anuales.

Para construir esta variable se calcularon las variaciones porcentuales del total de ventas y del total de compras respecto al año inmediatamente anterior. El año 2017 no cuenta con variación interanual debido a la ausencia de un período previo de comparación, mientras que el año 2026 se mantiene como referencia parcial y debe interpretarse con cautela.

Tabla 19

Variación interanual de compras y ventas declaradas, período 2017–2025

Año	Variación ventas (%)	Variación compras (%)
2017	N/D	N/D
2018	5,94	6,69
2019	1,77	-0,01
2020	-15,03	-17,28
2021	22,48	23,86
2022	14,61	13,85
2023	3,98	4,00
2024	1,94	1,81
2025	9,46	12,60

Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). La variación interanual corresponde al cambio porcentual del total de ventas y del total de compras respecto al año inmediatamente anterior. Para el año 2017 no se calcula variación debido a la ausencia de un período previo de comparación. El año 2026 fue excluido por corresponder a información parcial.

Los resultados de la Tabla 19 muestran que las compras y ventas declaradas presentan una dinámica similar a lo largo del período 2017–2025. En 2018 se observa un crecimiento moderado tanto en ventas como en compras, mientras que en 2019 las ventas continúan creciendo levemente y las compras se mantienen prácticamente sin variación.

El año 2020 registra la mayor contracción de la serie, con una reducción de 15,03 % en las ventas declaradas y de 17,28 % en las compras declaradas. Este comportamiento evidencia una caída significativa de la actividad económica declarada, coherente con la interrupción observada en otros indicadores durante ese período.

A partir de 2021 se observa una recuperación importante, con incrementos de 22,48 % en ventas y 23,86 % en compras. Posteriormente, el crecimiento se mantiene positivo, aunque

con menor intensidad en 2023 y 2024. En 2025 se registra nuevamente un aumento relevante, especialmente en las compras declaradas. En conjunto, este indicador permite identificar períodos de contracción y recuperación que serán útiles para el análisis exploratorio posterior y para la interpretación de posibles cambios en el comportamiento económico declarado.

5.3.3 Balance declarado y relación compras/ventas

El balance declarado y el ratio compras/ventas constituyen variables derivadas orientadas a analizar la relación entre los valores de ventas y compras reportados por los contribuyentes. Estas métricas permiten observar, a nivel agregado, si los ingresos declarados superan a los egresos registrados por compras, así como evaluar la proporción que representan las compras frente al total de ventas en cada período.

Para construir estas variables se utilizó la base homologada de compras y ventas declaradas. El balance declarado fue calculado como la diferencia entre el total de ventas y el total de compras, mientras que el ratio compras/ventas se obtuvo dividiendo el total de compras entre el total de ventas. Estos indicadores permiten complementar el análisis de los valores absolutos, ya que aportan una medida relativa de comportamiento económico declarado.

Tabla 20
Balance declarado y ratio compras/ventas, periodo 2017–2026

Año	Balance declarado	Ratio compras/ventas
2017	31,80	0,82
2018	32,58	0,83
2019	35,95	0,81
2020	34,06	0,79
2021	39,94	0,80
2022	47,00	0,79
2023	48,82	0,80
2024	50,00	0,79
2025	48,67	0,82
2026	7,09	0,83

Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). El balance declarado corresponde a la diferencia entre el total de ventas y el total de compras, expresado en miles de millones de dólares estadounidenses. El ratio compras/ventas se calculó dividiendo el total de compras entre el total de ventas. La información correspondiente al año 2026 es parcial y debe interpretarse con cautela.

Los resultados de la Tabla 20 muestran que el balance declarado se mantiene positivo durante todos los años analizados, lo que indica que, a nivel agregado, las ventas declaradas superan a las compras declaradas en cada período. Entre 2017 y 2024 se observa un crecimiento progresivo del balance, pasando de 31,80 a 50,00 mil millones de dólares, mientras que en 2025 se registra una ligera reducción a 48,67 mil millones.

El ratio compras/ventas presenta una alta estabilidad durante la serie, con valores comprendidos entre 0,79 y 0,83. Esto significa que, en términos agregados, las compras declaradas representan aproximadamente entre el 79 % y el 83 % de las ventas declaradas. Esta estabilidad aporta una referencia útil para análisis posteriores, ya que desviaciones importantes respecto a este comportamiento podrían constituir señales de interés analítico.

El año 2026 mantiene un balance positivo y un ratio similar al de los años anteriores; sin embargo, su interpretación debe realizarse con cautela debido a que corresponde a información parcial. Por esta razón, las comparaciones principales deberán concentrarse en el período 2017–2025.

5.3.4 Participación provincial en ventas declaradas

La participación provincial en ventas declaradas permite identificar el peso relativo de cada territorio dentro del total de actividad económica reportada. Este indicador resulta relevante porque permite comparar la concentración territorial de las ventas con la concentración observada previamente en la recaudación tributaria, aportando una perspectiva complementaria sobre la distribución espacial de la actividad económica declarada.

Para construir esta variable se agrupó el total de ventas por provincia y año. Posteriormente, se calculó la participación porcentual de cada provincia respecto al total nacional de ventas declaradas en cada período. Este procedimiento permite identificar las provincias con mayor concentración de ventas y observar si dicha distribución se mantiene estable o presenta cambios a lo largo del tiempo.

Tabla 21*Participación provincial en ventas declaradas, principales provincias, período 2017–2025*

Provincia	2017	2018	2019	2020	2021	2022	2023	2024	2025	Total
Pichincha	42,98	42,09	41,53	38,60	39,14	39,37	39,15	38,34	36,41	39,53
Guayas	31,75	32,32	32,97	35,08	34,80	34,69	35,01	35,08	35,67	34,28
Azuay	4,76	4,83	4,74	4,71	4,72	4,60	3,97	3,88	4,03	4,42
Manabí	4,21	4,13	3,98	4,26	4,14	4,18	4,24	4,35	4,60	4,25
El Oro	3,01	3,06	3,13	3,27	3,32	3,21	3,16	2,97	3,13	3,14
Tungurahua	2,41	2,46	2,43	2,18	2,20	2,25	2,21	2,09	2,13	2,25
Los Ríos	1,71	1,77	1,80	2,20	1,86	1,78	1,80	2,26	2,31	1,96
Santo Domingo de los Tsáchilas	1,46	1,46	1,48	1,57	1,52	1,49	1,55	1,75	1,87	1,59
Esmeraldas	1,02	1,00	0,94	1,04	0,89	0,96	1,10	1,18	1,16	1,04
Cotopaxi	0,97	0,99	1,00	1,04	0,98	0,94	0,98	1,00	1,04	0,99

Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el

Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual de cada provincia

respecto al total anual de ventas declaradas. La columna Total corresponde a la participación acumulada de cada provincia sobre el total de ventas declaradas durante el período 2017–2025. Se presentan las diez provincias con mayor participación acumulada. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

Los resultados de la Tabla 21 evidencian una concentración territorial importante de las ventas declaradas en las provincias de Pichincha y Guayas. Pichincha presenta la mayor participación acumulada del período 2017–2025, con 39,53 %, mientras que Guayas alcanza 34,28 %. En conjunto, ambas provincias concentran más del 70 % de las ventas declaradas, lo que refleja el peso económico de los principales polos productivos y comerciales del país.

A diferencia de lo observado en la recaudación tributaria, donde Pichincha concentra una proporción superior al 50 %, en las ventas declaradas su participación es menor y muestra una tendencia descendente, pasando de 42,98 % en 2017 a 36,41 % en 2025. En contraste, Guayas incrementa gradualmente su participación, alcanzando 35,67 % en 2025. Este comportamiento sugiere una distribución territorial de la actividad económica declarada algo menos concentrada que la recaudación tributaria.

Las demás provincias presentan participaciones considerablemente menores. Azuay, Manabí y El Oro se ubican como los siguientes territorios con mayor peso relativo, aunque con valores acumulados inferiores al 5 %. Estos resultados permiten construir una referencia territorial útil para comparar la concentración económica declarada frente a la concentración de la recaudación tributaria en los análisis posteriores.

5.3.5 Participación sectorial en ventas declaradas

La participación sectorial en ventas declaradas permite identificar qué sectores económicos concentran una mayor proporción de la actividad económica reportada. Este indicador resulta relevante porque el comportamiento de las ventas no se distribuye de manera uniforme entre sectores, sino que tiende a concentrarse en determinadas actividades productivas, comerciales o de servicios.

Para construir esta variable se agrupó el total de ventas declaradas por año y código de sector económico. Posteriormente, se calculó la participación porcentual de cada sector respecto al total anual de ventas declaradas. Este procedimiento permite observar la importancia relativa de cada sector dentro de la actividad económica reportada y analizar si dicha composición se mantiene estable o presenta cambios durante el período estudiado.

Tabla 22

Participación sectorial en ventas declaradas, principales sectores, período 2017–2025

Código de sector	2017	2018	2019	2020	2021	2022	2023	2024	2025	Total
G	42,99	44,27	44,55	44,58	45,10	45,60	44,71	45,08	44,67	44,67
C	15,91	15,85	15,03	15,50	15,54	15,44	14,84	14,29	14,18	15,10
A	7,33	7,39	7,79	8,83	8,79	8,63	8,29	9,13	9,69	8,50
H	4,76	4,82	4,83	4,37	4,49	4,74	5,18	5,02	5,11	4,84
M	4,65	4,40	4,22	3,91	3,88	3,80	4,02	3,88	3,88	4,05
B	2,63	2,48	2,70	2,82	3,25	3,21	3,50	3,42	3,41	3,09
F	4,02	3,23	2,79	2,28	2,33	2,39	2,55	2,45	2,50	2,69
K	2,55	2,54	2,66	2,88	2,56	2,67	2,72	2,81	2,67	2,67
J	3,01	2,95	2,97	2,90	2,48	2,22	2,19	2,10	1,99	2,48
Q	1,75	1,81	1,98	2,27	2,37	2,12	2,19	1,95	2,02	2,05

Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual de cada código de sector económico respecto al total anual de ventas declaradas. La columna Total corresponde a la participación acumulada de cada sector sobre el total de ventas declaradas durante el período 2017–2025. Se presentan los diez sectores con mayor participación acumulada. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

Significado de sectores económicos

G = Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas

C = Industrias manufactureras

A = Agricultura, ganadería, silvicultura y pesca

H = Transporte y almacenamiento

M = Actividades profesionales, científicas y técnicas

B = Explotación de minas y canteras

F = Construcción

K = Actividades financieras y de seguros

J = Información y comunicación

Q = Actividades de atención de la salud humana y de asistencia social

Los resultados de la Tabla 22 muestran una alta concentración sectorial de las ventas declaradas. El sector G, correspondiente al comercio al por mayor y al por menor, reparación de vehículos automotores y motocicletas, presenta la mayor participación acumulada del período 2017–2025, con 44,67 %. Este resultado evidencia el peso predominante de las actividades comerciales dentro de la actividad económica declarada.

En segundo lugar, se ubica el sector C, correspondiente a industrias manufactureras, con una participación acumulada de 15,10 %. Aunque mantiene un peso relevante dentro de la estructura económica, se observa una ligera disminución hacia los últimos años del período,

pasando de 15,91 % en 2017 a 14,18 % en 2025. Por su parte, el sector A, asociado a agricultura, ganadería, silvicultura y pesca, ocupa el tercer lugar con 8,50 % y presenta una tendencia creciente, alcanzando 9,69 % en 2025.

Los demás sectores presentan participaciones menores, aunque relevantes para caracterizar la composición económica declarada. Entre ellos se encuentran transporte y almacenamiento, actividades profesionales, explotación de minas y canteras, construcción, actividades financieras, información y comunicación, y salud humana y asistencia social. En conjunto, estos resultados permiten identificar una estructura de ventas fuertemente concentrada en comercio, manufactura y actividades agropecuarias, lo cual será relevante para comparar posteriormente la concentración económica declarada frente a la recaudación tributaria y construir perfiles sectoriales diferenciados.

5.3.6 Comparación sectorial entre ventas y compras declaradas

La participación sectorial en compras declaradas permite identificar qué sectores económicos concentran una mayor proporción de los valores reportados como compras dentro de la información homologada. Este indicador complementa el análisis de ventas declaradas, ya que permite observar si los sectores con mayor nivel de ingresos declarados también concentran una proporción similar de compras.

Para construir esta variable se agrupó el total de compras declaradas por año y código de sector económico. Posteriormente, se calculó la participación porcentual de cada sector respecto al total anual de compras declaradas. Este procedimiento permite comparar la estructura sectorial de compras con la estructura sectorial de ventas y aporta una base para interpretar relaciones económicas agregadas entre ingresos y egresos declarados.

Tabla 23*Comparación sectorial entre ventas y compras declaradas, principales sectores, período 2017–2025*

Código de sector	Sector económico	Participación en ventas (%)	Participación en compras (%)	Diferencia ventas-compras
G	Comercio al por mayor y menor; reparación de vehículos	44,67	44,25	0,42
C	Industrias manufactureras	15,10	14,71	0,39
A	Agricultura, ganadería, silvicultura y pesca	8,50	8,12	0,38
H	Transporte y almacenamiento	4,84	4,28	0,56
M	Actividades profesionales, científicas y técnicas	4,05	3,22	0,83
B	Explotación de minas y canteras	3,09	2,95	0,14
F	Construcción	2,69	2,66	0,03
K	Actividades financieras y de seguros	2,67	3,14	-0,47
Q	Actividades de atención de la salud humana y asistencia social	2,05	2,14	-0,09

Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual acumulada de cada sector económico sobre el total de ventas y compras declaradas durante el período 2017–2025. La diferencia ventas-compras se expresa en puntos porcentuales y corresponde a la resta entre la participación en ventas y la participación en compras. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

Los resultados de la Tabla 23 muestran que los principales sectores económicos presentan una relación relativamente equilibrada entre su participación en ventas y compras declaradas. El sector G, correspondiente al comercio al por mayor y menor y reparación de vehículos, concentra la mayor participación tanto en ventas como en compras, con valores superiores al 44 % en ambas dimensiones. Este comportamiento confirma el peso predominante del comercio dentro de la actividad económica declarada.

Los sectores C, A y H también mantienen participaciones similares entre ventas y compras, lo que sugiere una estructura proporcional entre ingresos y egresos declarados a nivel

agregado. En cambio, el sector M presenta una participación en ventas ligeramente superior a la de compras, mientras que el sector K muestra una participación en compras mayor que en ventas.

En conjunto, esta comparación permite identificar sectores con estructuras relativamente equilibradas y otros con diferencias moderadas entre ventas y compras. Estas diferencias no constituyen evidencia de riesgo por sí mismas, pero aportan una referencia útil para posteriores análisis sectoriales, especialmente al construir indicadores de balance económico, ratios compras/ventas y posibles perfiles de comportamiento atípico.

5.3.7 Síntesis de variables derivadas de compras y ventas declaradas

Las variables derivadas de compras y ventas declaradas permitieron complementar el análisis de recaudación tributaria mediante indicadores asociados directamente a la actividad económica reportada. A partir de la base homologada, se construyeron métricas de ventas, compras, exportaciones, importaciones, balance declarado y ratio compras/ventas, las cuales aportan una visión agregada sobre el comportamiento económico de los contribuyentes durante el período analizado.

Los resultados muestran una tendencia general de crecimiento en las ventas y compras declaradas entre 2017 y 2025, con una contracción importante en 2020 y una recuperación sostenida a partir de 2021. Este comportamiento es consistente con la dinámica observada previamente en la recaudación tributaria, lo que refuerza la importancia de considerar la dimensión temporal dentro de los análisis posteriores.

Desde la perspectiva territorial, las ventas declaradas se concentran principalmente en Pichincha y Guayas, aunque con una distribución menos concentrada que la observada en la recaudación tributaria. Este hallazgo sugiere que la actividad económica declarada y la recaudación no se distribuyen exactamente de la misma forma en el territorio, por lo que ambas dimensiones deben analizarse de manera complementaria (Censos, s. f.).

En términos sectoriales, se identificó una alta concentración de ventas y compras en el comercio, seguido por industrias manufactureras y actividades agropecuarias. La comparación entre la participación sectorial en ventas y compras permitió observar relaciones relativamente equilibradas en los principales sectores, aunque también se identificaron diferencias moderadas que podrían ser relevantes para la construcción de indicadores de balance económico y perfiles de comportamiento sectorial.

En conjunto, las variables derivadas de compras y ventas declaradas constituyen un insumo relevante para las etapas posteriores de análisis exploratorio y modelado. Estas métricas permiten aproximarse a la dinámica económica declarada, comparar comportamientos territoriales y sectoriales, y construir indicadores útiles para la identificación de patrones atípicos asociados al riesgo tributario.

CAPÍTULO 6

ANÁLISIS EXPLORATORIO DE DATOS

El presente capítulo desarrolla el análisis exploratorio de los datos utilizados en la investigación, a partir de las variables analíticas construidas en el capítulo anterior. Esta etapa tiene como finalidad examinar visual y estadísticamente el comportamiento de los principales indicadores tributarios y económicos, identificando tendencias, patrones de concentración, variaciones temporales y posibles comportamientos atípicos.

El análisis exploratorio constituye una fase fundamental dentro del proceso de ciencia de datos, ya que permite comprender la estructura de la información antes de aplicar modelos analíticos o técnicas de aprendizaje no supervisado. A través de gráficos, tablas resumen y medidas descriptivas, se busca reconocer comportamientos relevantes que orienten la posterior segmentación e identificación de perfiles de interés tributario (*What is Exploratory Data Analysis?* | IBM, s. f.).

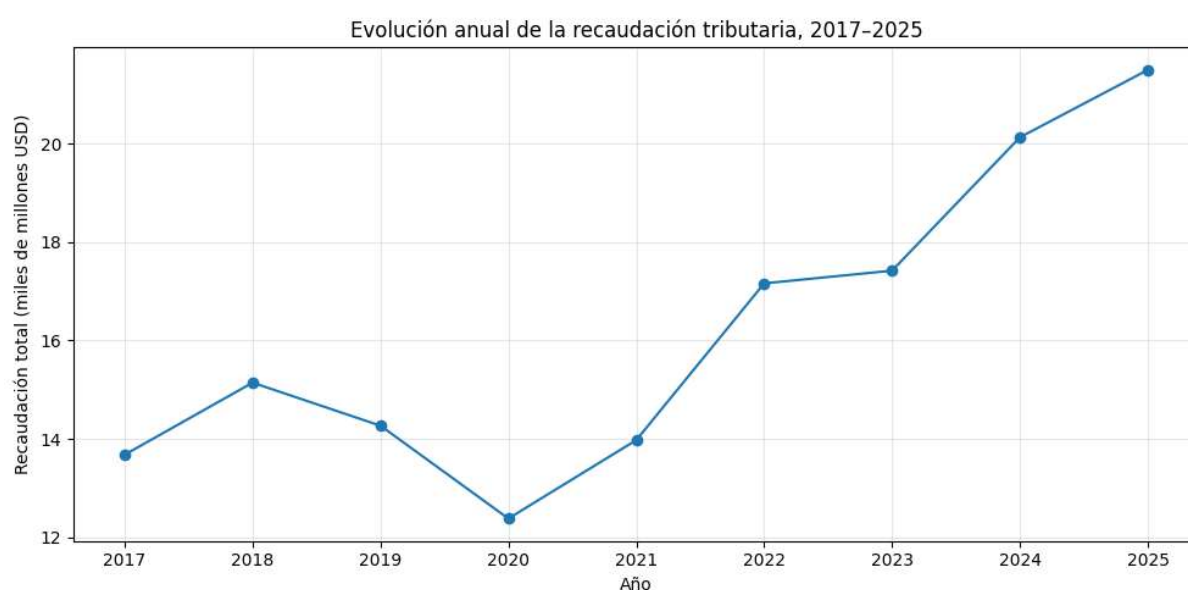
En este capítulo se analizan principalmente patrones temporales, distribución territorial de la recaudación, concentración por tipo de impuesto y diferencias según tipo de contribuyente. Los resultados obtenidos no deben interpretarse como evidencia concluyente de evasión fiscal, sino como señales descriptivas iniciales que permiten orientar el análisis posterior hacia segmentos, períodos o territorios con comportamientos diferenciados.

6.1 Análisis exploratorio de la recaudación tributaria

El análisis exploratorio de la recaudación tributaria permite examinar el comportamiento de los ingresos fiscales desde una perspectiva temporal, territorial y composicional. A partir de las variables derivadas construidas en el capítulo anterior, se revisan tendencias anuales, variaciones interanuales, patrones mensuales, concentración provincial y distribución por grupos de impuesto.

Esta primera aproximación exploratoria permite identificar regularidades y cambios relevantes dentro de la serie histórica, así como posibles desviaciones que podrían ser analizadas con mayor profundidad en las etapas posteriores de modelado. La recaudación tributaria constituye una fuente central dentro del estudio debido a que refleja, de manera agregada, la interacción entre actividad económica, obligaciones fiscales y comportamiento de los distintos tipos de contribuyentes.

Figura 2 Evolución anual de la recaudación tributaria total, período 2017–2025



Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). Los valores se expresan en miles de millones de dólares estadounidenses. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

La Figura 2 muestra la evolución anual de la recaudación tributaria total durante el período 2017–2025. Se observa un crecimiento inicial entre 2017 y 2018, seguido de una reducción en 2019 y una caída más pronunciada en 2020, año que registra el menor valor de la serie analizada. Este comportamiento refleja una interrupción importante en la tendencia de crecimiento observada al inicio del período.

A partir de 2021 se evidencia una recuperación progresiva de la recaudación tributaria, con un incremento significativo en 2022 y una tendencia ascendente sostenida hasta 2025. Los

años 2024 y 2025 presentan los valores más altos de la serie, lo que sugiere una recuperación y expansión de los ingresos tributarios posteriores a la contracción observada en 2020.

Este comportamiento temporal permite identificar tres momentos relevantes: una etapa inicial de crecimiento moderado, una fase de contracción asociada al año 2020 y una etapa posterior de recuperación sostenida. Estos patrones serán considerados en los análisis posteriores, especialmente al evaluar variaciones interanuales, posibles rupturas de tendencia y comportamientos atípicos dentro de la serie histórica.

Como complemento al análisis de la evolución anual en valores absolutos, se calculó la variación interanual de la recaudación tributaria. Este indicador permite observar con mayor claridad los cambios porcentuales registrados entre un año y otro, facilitando la identificación de períodos de crecimiento, contracción y recuperación dentro de la serie histórica.

Figura 3 Variación interanual de la recaudación tributaria, período 2018–2025



Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). La variación interanual corresponde al cambio porcentual de la recaudación total anual respecto al año inmediatamente anterior. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

La Figura 3 permite observar con mayor claridad los cambios porcentuales en la recaudación tributaria a lo largo del período 2018–2025. En 2018 se registra un crecimiento positivo respecto al año anterior, mientras que en 2019 se evidencia una contracción moderada.

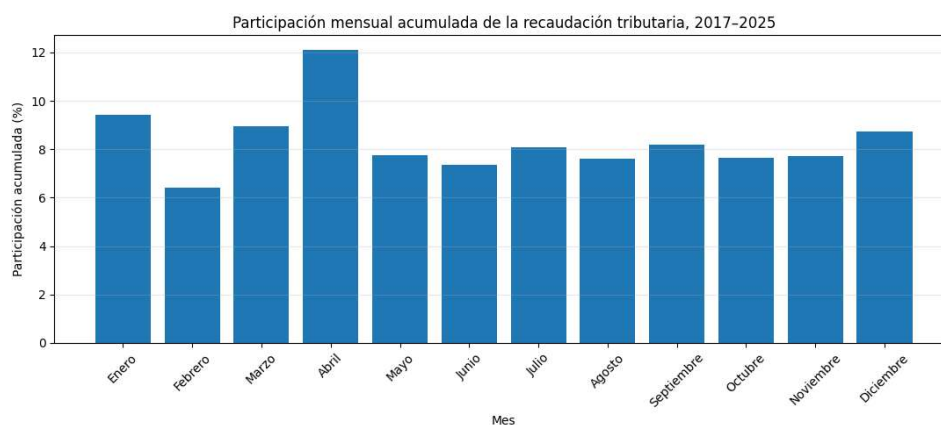
La reducción se profundiza en 2020, año que presenta la mayor caída interanual de la serie analizada, con una variación negativa cercana al 13 %.

A partir de 2021 se observa una recuperación significativa de la recaudación tributaria, con variaciones positivas en todos los años posteriores. El mayor crecimiento se registra en 2022, con una variación superior al 20 %, lo que refleja una recuperación importante después de la contracción observada en 2020. En 2023 el crecimiento se modera, mientras que en 2024 vuelve a presentarse un incremento relevante y en 2025 la variación positiva se mantiene, aunque con menor intensidad.

Este comportamiento confirma la existencia de una ruptura temporal en 2020 y una fase posterior de recuperación sostenida. Desde una perspectiva exploratoria, estos cambios interanuales resultan relevantes porque permiten identificar períodos que podrían influir en la construcción de variables temporales, en la detección de valores atípicos y en la interpretación de patrones de recaudación durante las etapas posteriores del análisis.

Luego de revisar la evolución anual y la variación interanual de la recaudación tributaria, se procedió a analizar su comportamiento mensual. Este análisis permite identificar si la recaudación presenta patrones estacionales o meses con mayor concentración relativa dentro del año, información relevante para interpretar variaciones temporales que podrían no observarse en una escala anual.

Figura 4 Participación mensual acumulada de la recaudación tributaria, período 2017–2025



Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual acumulada de cada mes respecto al total recaudado durante el período 2017–2025. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

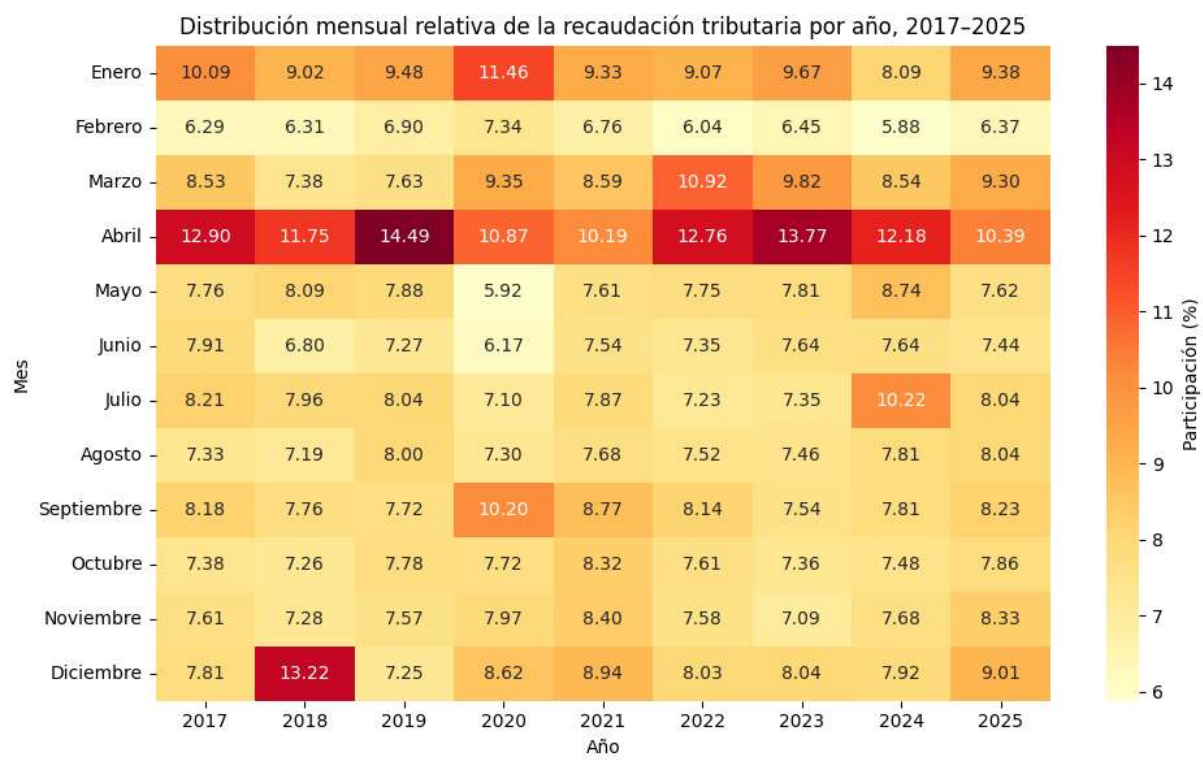
La Figura 4 muestra la participación mensual acumulada de la recaudación tributaria durante el período 2017–2025. Se observa que abril concentra la mayor proporción del total recaudado, con un valor superior al 12 %, lo que lo convierte en el mes de mayor peso relativo dentro de la estructura anual. También destacan enero, marzo y diciembre, cuyos porcentajes se ubican por encima del promedio mensual esperado, evidenciando períodos de mayor concentración recaudatoria.

En contraste, febrero presenta la menor participación acumulada del período, seguido de junio y agosto, meses que registran un peso relativo más bajo dentro del total observado. En términos generales, la distribución mensual no es completamente homogénea, ya que ciertos meses concentran una mayor proporción de ingresos tributarios, lo que sugiere la existencia de patrones estacionales en la dinámica de recaudación.

Estos resultados permiten identificar una estructura temporal recurrente en la recaudación tributaria, donde algunos meses presentan una mayor importancia relativa dentro del año fiscal. Este comportamiento será útil en los análisis posteriores, especialmente al comparar variaciones mensuales, detectar posibles anomalías y contextualizar cambios interanuales dentro de la serie histórica.

Con el fin de profundizar en el comportamiento temporal de la recaudación tributaria, se analizó la distribución relativa de los meses dentro de cada año. Esta visualización permite identificar patrones estacionales recurrentes, así como cambios en la intensidad de la recaudación mensual entre distintos períodos de la serie histórica.

Figura 5 Distribución mensual relativa de la recaudación tributaria por año, período 2017–2025



Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). Los valores expresan la participación porcentual de cada mes respecto al total recaudado en su correspondiente año. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

La Figura 5 permite observar la distribución mensual relativa de la recaudación tributaria dentro de cada año del período 2017–2025. En términos generales, se identifica un patrón relativamente estable a lo largo de la serie, con mayor concentración de recaudación en abril, mes que registra los porcentajes más elevados en la mayoría de los años analizados. Este comportamiento sugiere la existencia de una estacionalidad recurrente en la estructura mensual de la recaudación.

Asimismo, se aprecia que febrero tiende a presentar una menor participación relativa en casi todos los años, mientras que meses como enero, marzo y diciembre suelen ubicarse por encima del promedio en varios períodos. Esta distribución evidencia que la recaudación

tributaria no se reparte de manera uniforme durante el año, sino que presenta concentraciones temporales específicas.

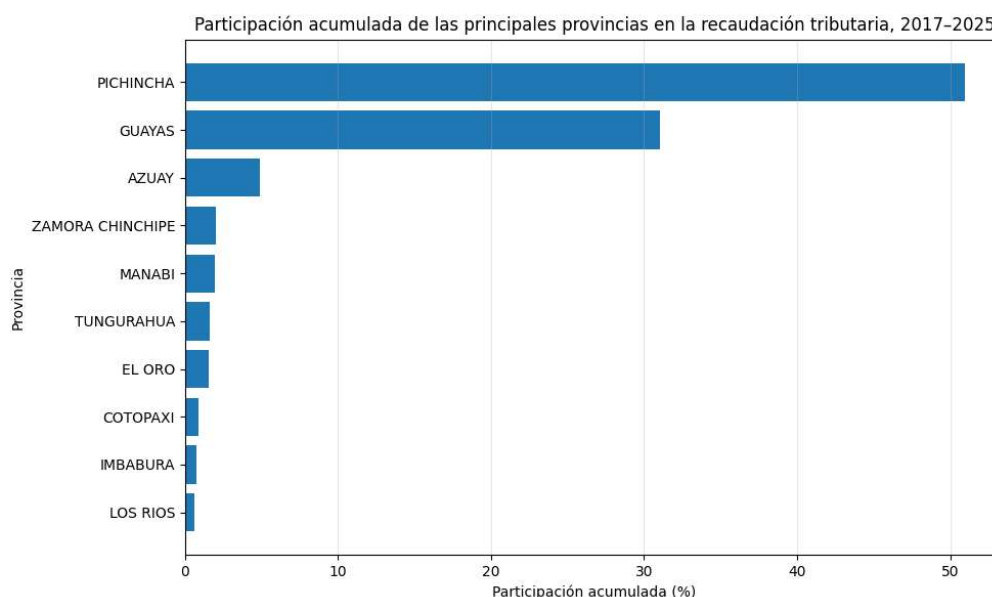
El mapa de calor también permite identificar ciertos comportamientos particulares. Por ejemplo, en 2018 diciembre muestra una participación relativamente alta, mientras que en 2024 destaca julio con un valor superior al observado en la mayoría de años previos. De igual forma, en 2020 se observa una distribución menos homogénea, con concentraciones relativamente más marcadas en enero, abril y septiembre, lo que puede asociarse a alteraciones coyunturales en la dinámica recaudatoria de ese período.

En conjunto, esta visualización confirma la presencia de patrones temporales recurrentes y facilita la identificación de meses con mayor o menor peso relativo dentro de la recaudación anual. Estos hallazgos constituyen una base útil para posteriores análisis comparativos, detección de anomalías y evaluación de regularidades estacionales en la serie histórica.

6.1.1 Concentración territorial de la recaudación tributaria

Después de analizar la evolución temporal de la recaudación tributaria, se procedió a examinar su distribución territorial. Este análisis permite identificar el grado de concentración geográfica de los ingresos fiscales y reconocer las provincias con mayor peso relativo dentro del total recaudado a nivel nacional.

La concentración territorial constituye un aspecto relevante dentro del análisis exploratorio, debido a que la actividad económica y tributaria no se distribuye de manera uniforme en el país. Por esta razón, observar la participación provincial permite reconocer territorios con alta incidencia fiscal, así como provincias con menor participación relativa, lo cual aporta insumos para la construcción de perfiles analíticos territoriales.

Figura 6 Participación acumulada de las principales provincias en la recaudación tributaria, período 2017–2025

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual acumulada de cada provincia respecto al total recaudado durante el período 2017–2025. Se presentan las diez provincias con mayor participación acumulada. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

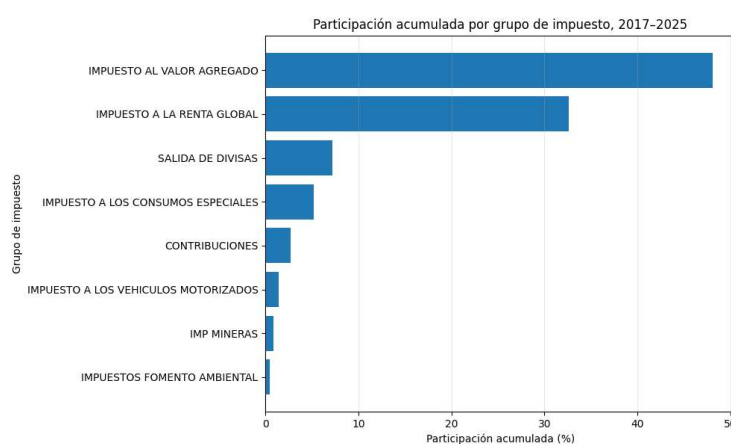
La Figura 6 evidencia una alta concentración territorial de la recaudación tributaria en un número reducido de provincias. Pichincha concentra la mayor participación acumulada del período 2017–2025, con aproximadamente 50,96 % del total recaudado, seguida por Guayas con 31,02 %. En conjunto, ambas provincias representan cerca del 81,98 % de la recaudación acumulada, lo que refleja una marcada centralización espacial de los ingresos tributarios.

En un segundo nivel de participación aparece Azuay, aunque con una proporción considerablemente menor, equivalente al 4,87 %. A partir de este punto, el resto de provincias incluidas en la figura presentan participaciones individuales reducidas, inferiores al 2 % en la mayoría de las provincias. Este patrón muestra que la estructura territorial de la recaudación se encuentra fuertemente dominada por los principales polos económicos y administrativos del país.

Desde una perspectiva exploratoria, estos resultados sugieren que el análisis tributario posterior debe considerar la existencia de una asimetría territorial significativa, ya que la dinámica observada a nivel nacional puede estar explicada, en gran medida, por el comportamiento de Pichincha y Guayas. Por tanto, la concentración geográfica constituye un rasgo clave para interpretar la distribución de la actividad económica y de la base recaudatoria en el Ecuador.

6.1.2 Composición de la recaudación por grupos de impuesto

Figura 7 Participación acumulada por grupo de impuesto en la recaudación tributaria, período 2017–2025



Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). La figura presenta la participación porcentual acumulada de los ocho principales grupos de impuesto dentro de la recaudación tributaria total del período 2017–2025. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

La Figura 7 muestra que la estructura de la recaudación tributaria en el período 2017–2025 se encuentra fuertemente concentrada en dos grupos de impuesto. En primer lugar, el Impuesto al Valor Agregado representa la mayor participación acumulada, con cerca de la mitad del total recaudado, lo que evidencia su papel central dentro del sistema tributario. En segundo lugar, el Impuesto a la Renta Global constituye el segundo componente más relevante, aportando alrededor de un tercio de la recaudación total.

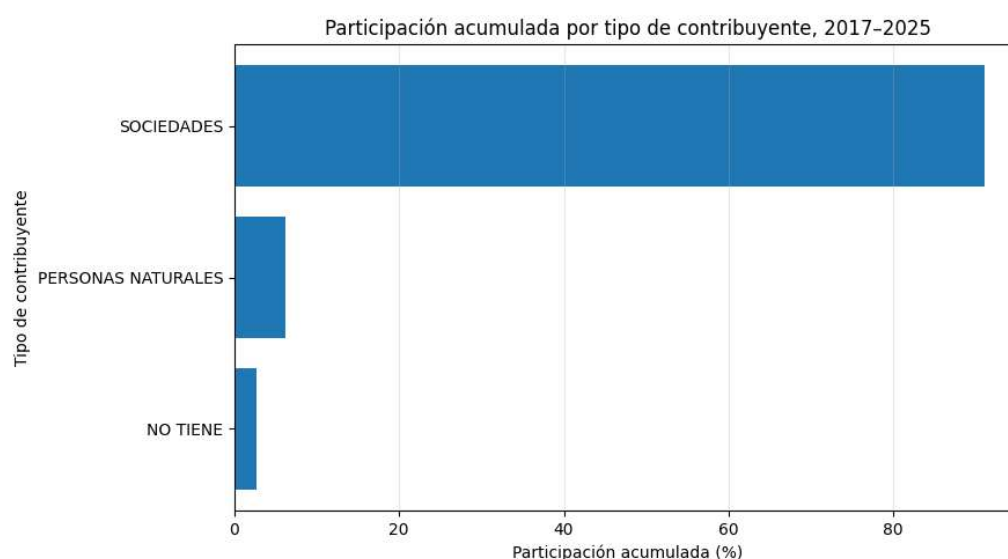
A considerable distancia de estos dos grupos principales se ubican la Salida de Divisas y el Impuesto a los Consumos Especiales, cuya participación es significativamente menor, aunque todavía relevante dentro de la estructura tributaria. En cambio, grupos como Contribuciones, Impuesto a los Vehículos Motorizados, Impuestos Mineros e Impuestos de Fomento Ambiental presentan pesos reducidos, lo que indica una contribución más limitada al total recaudado.

En conjunto, estos resultados evidencian una estructura tributaria altamente concentrada en pocos componentes, especialmente en tributos asociados al consumo y a la renta. Este patrón resulta importante para el análisis posterior, ya que permite identificar cuáles son las fuentes predominantes de ingresos fiscales y cómo su comportamiento puede influir en la dinámica global de la recaudación.

6.1.3 Composición de la recaudación según tipo de contribuyente

Luego de analizar la composición de la recaudación por grupo de impuesto, se examinó su distribución según el tipo de contribuyente. Este análisis permite identificar el peso relativo de sociedades, personas naturales y registros sin clasificación específica dentro del total recaudado, aportando una perspectiva complementaria sobre la estructura de la base tributaria.

Figura 8 Participación acumulada por tipo de contribuyente en la recaudación tributaria, período 2017–2025



Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). La figura presenta la participación porcentual acumulada de la recaudación tributaria según el tipo de contribuyente para el período 2017–2025. Se excluye el año 2026 debido a que la información disponible corresponde únicamente a un período parcial.

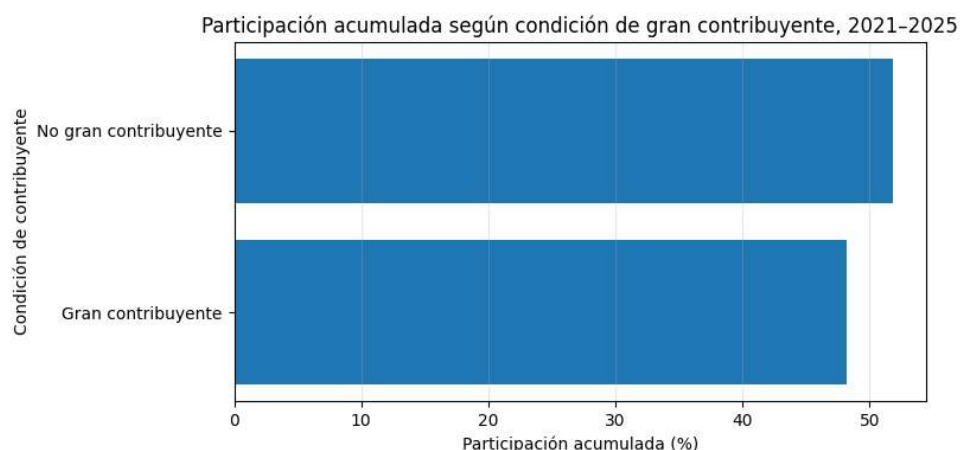
La Figura 8 muestra una marcada concentración de la recaudación tributaria en el segmento de sociedades, que acumula aproximadamente el 91.02 % del total recaudado durante el período analizado. En contraste, las personas naturales representan cerca del 6.25 %, mientras que la categoría “No tiene” concentra alrededor del 2.73 %.

Estos resultados evidencian que la estructura de la recaudación tributaria en Ecuador depende principalmente de los contribuyentes clasificados como sociedades, lo que sugiere una mayor contribución relativa de organizaciones empresariales dentro del total de ingresos fiscales. Por su parte, la participación de personas naturales es considerablemente menor, aunque sigue constituyendo un componente identificable dentro de la estructura tributaria.

La reducida proporción de registros agrupados en la categoría “No tiene” indica que, si bien existe una fracción de información sin clasificación específica, su peso dentro del total recaudado es limitado. En conjunto, este comportamiento confirma una alta concentración de la recaudación en un tipo particular de contribuyente, aspecto que deberá considerarse en etapas posteriores de segmentación y análisis del riesgo tributario.

6.1.4 Composición de la recaudación según condición de gran contribuyente

Finalmente, se analizó la participación de la recaudación tributaria según la condición de gran contribuyente. Esta dimensión permite evaluar el grado de concentración fiscal asociado a contribuyentes de mayor relevancia económica y tributaria, diferenciándolos del resto de contribuyentes registrados en la información consolidada.

Figura 9 Participación acumulada según condición de gran contribuyente, período 2021–2025

Nota. Elaboración propia a partir de los registros consolidados de recaudación tributaria publicados por el Servicio de Rentas Internas (SRI). La variable original fue normalizada en dos categorías: “Gran contribuyente” y “No gran contribuyente”. La primera agrupa los valores SI, G, Q y S, mientras que la segunda corresponde al valor NO. Los resultados se expresan como porcentaje de participación acumulada sobre la recaudación total del período 2021–2025.

La Figura 9 muestra una distribución relativamente equilibrada de la recaudación tributaria acumulada entre contribuyentes clasificados como grandes y no grandes durante el período 2021–2025. Se observa que los no grandes contribuyentes concentran una proporción ligeramente mayor, con aproximadamente 51,84 % del total recaudado, mientras que los grandes contribuyentes representan cerca del 48,16 %.

Este resultado sugiere que, aunque la recaudación se encuentra fuertemente concentrada en segmentos empresariales de alta relevancia económica, el conjunto de contribuyentes no clasificados como grandes mantiene también un peso significativo dentro de la estructura tributaria agregada. En otras palabras, la recaudación nacional no depende exclusivamente de un grupo reducido de grandes aportantes, sino que refleja una contribución importante del resto del universo tributario.

Desde una perspectiva exploratoria, esta evidencia justifica incorporar la condición de gran contribuyente como una dimensión analítica relevante en las etapas posteriores del

estudio, ya que permite diferenciar perfiles de recaudación y posibles patrones de comportamiento tributario entre segmentos con distinto peso económico y fiscal.

6.1.5 Síntesis del análisis exploratorio de la recaudación tributaria

El análisis exploratorio de la recaudación tributaria permitió identificar patrones relevantes en la evolución temporal, distribución territorial, composición tributaria y estructura de contribuyentes durante el período analizado. En términos temporales, se observó una caída importante de la recaudación en 2020, seguida de una recuperación sostenida a partir de 2021 y un crecimiento progresivo hasta 2025. Este comportamiento evidencia una ruptura en la serie histórica y una posterior fase de expansión de los ingresos tributarios.

El análisis mensual mostró que la recaudación no se distribuye de manera uniforme a lo largo del año. Abril concentra la mayor participación relativa, mientras que febrero presenta el menor peso acumulado. Asimismo, el mapa de calor permitió observar patrones estacionales recurrentes y algunos comportamientos particulares en años específicos, lo que confirma la utilidad de incorporar variables temporales dentro de los análisis posteriores.

Desde la perspectiva territorial, se evidenció una alta concentración geográfica de la recaudación tributaria. Pichincha y Guayas concentran la mayor parte del total recaudado, lo que refleja el peso económico y fiscal de los principales polos productivos y administrativos del país. Esta concentración territorial constituye un elemento relevante para la construcción de perfiles analíticos, ya que las dinámicas nacionales pueden estar fuertemente influenciadas por el comportamiento de estas provincias.

En cuanto a la composición de la recaudación, se identificó que el Impuesto al Valor Agregado y el Impuesto a la Renta Global constituyen los principales componentes tributarios, concentrando la mayor parte de los ingresos fiscales. De igual manera, el análisis por tipo de contribuyente evidenció que las sociedades representan el segmento con mayor peso dentro de la recaudación acumulada, mientras que la condición de gran contribuyente permitió reconocer

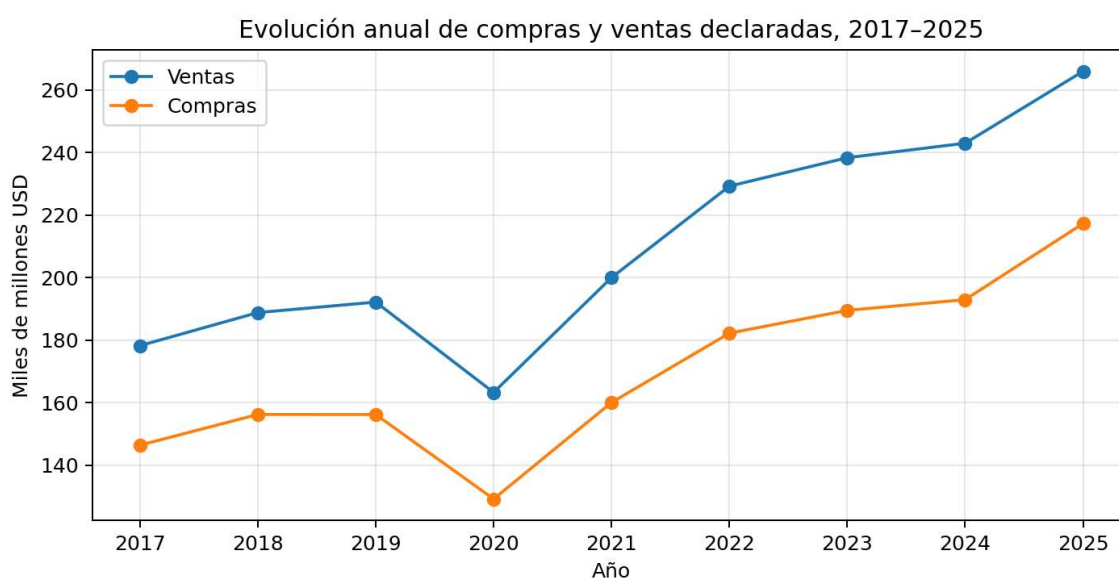
una distribución relativamente equilibrada entre grandes y no grandes contribuyentes durante el período 2021–2025.

En conjunto, los hallazgos exploratorios muestran que la recaudación tributaria presenta patrones de concentración temporal, territorial y estructural que deben ser considerados en las etapas posteriores de la investigación. Estos resultados no constituyen evidencia concluyente de riesgo tributario o evasión fiscal, pero permiten identificar dimensiones relevantes para la construcción de variables, la segmentación de perfiles y la posterior aplicación de modelos de aprendizaje no supervisado.

6.2 Análisis exploratorio de compras y ventas declaradas

El análisis exploratorio de compras y ventas declaradas complementa la revisión de la recaudación tributaria, ya que permite observar la dinámica económica reportada por los contribuyentes desde la perspectiva de la actividad comercial agregada. En esta sección se examinan la evolución anual de ventas y compras, sus variaciones interanuales y la distribución territorial y sectorial de las ventas declaradas.

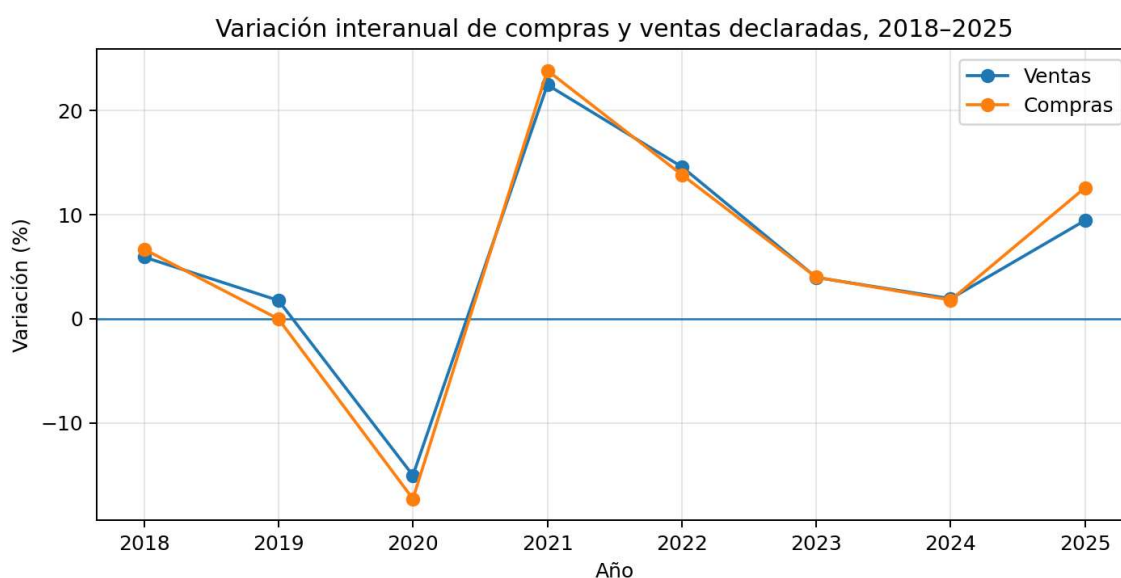
Figura 10 Evolución anual de compras y ventas declaradas, período 2017–2025



Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). Los valores se expresan en miles de millones de dólares estadounidenses. Se excluye el año 2026 debido a que corresponde a información parcial.

La Figura 10 muestra que las ventas y compras declaradas mantienen una trayectoria similar durante el período 2017–2025. Ambas series presentan una reducción importante en 2020 y una recuperación sostenida a partir de 2021. Esta relación evidencia que la dinámica económica declarada siguió un patrón consistente con la evolución general observada en la recaudación tributaria.

Figura 11 Variación interanual de compras y ventas declaradas, período 2018–2025



Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). La variación corresponde al cambio porcentual respecto al año inmediatamente anterior.

La Figura 10 permite observar con mayor claridad la contracción registrada en 2020 y la recuperación posterior en 2021. Las compras y ventas presentan variaciones interanuales similares, lo que sugiere una relación agregada relativamente estable entre ambas dimensiones. Estos cambios son relevantes para el análisis posterior, ya que permiten contextualizar posibles comportamientos atípicos dentro de períodos de expansión o contracción económica.

Figura 12 Participación acumulada de las principales provincias en ventas declaradas, período 2017–2025

Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). Se presentan las diez provincias con mayor participación acumulada. Se excluye el año 2026 debido a su cobertura parcial.

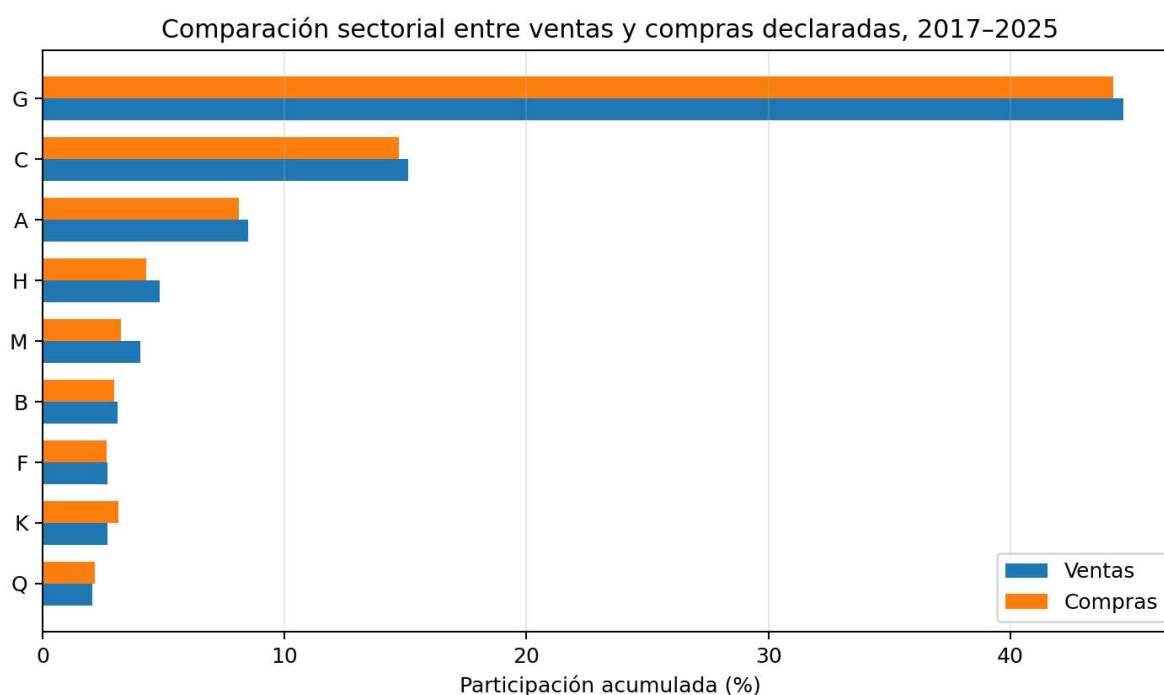
La Figura 12 evidencia que las ventas declaradas se concentran principalmente en Pichincha y Guayas. Sin embargo, esta concentración es menor que la observada en la recaudación tributaria, lo que sugiere que la actividad económica declarada y la recaudación no se distribuyen exactamente de la misma forma en el territorio. Esta diferencia resulta útil para interpretar brechas o comportamientos diferenciados entre actividad económica y aporte tributario.

Figura 13 Participación acumulada por sector económico en ventas declaradas, período 2017–2025

Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). Los sectores se identifican mediante los códigos de actividad económica utilizados en la fuente original.

La Figura 13 muestra una concentración sectorial importante de las ventas declaradas en el sector G, correspondiente al comercio al por mayor y menor, seguido por industrias manufactureras y actividades agropecuarias. Este resultado confirma que la actividad económica declarada se concentra en un número reducido de sectores, aspecto que debe considerarse al construir perfiles analíticos y comparar patrones de comportamiento tributario.

Figura 14 Comparación sectorial entre ventas y compras declaradas, período 2017–2025



Nota. Elaboración propia a partir de la base homologada de compras y ventas declaradas publicada por el Servicio de Rentas Internas (SRI). Los valores representan la participación porcentual acumulada de cada sector sobre el total de ventas y compras declaradas durante el período 2017–2025.

La Figura 14 permite comparar la participación acumulada de ventas y compras en los principales sectores económicos. En general, los sectores con mayor participación en ventas también presentan una participación elevada en compras, lo que refleja una estructura agregada relativamente equilibrada. No obstante, las diferencias moderadas entre ambas variables

aportan una referencia útil para el análisis posterior de perfiles sectoriales y posibles comportamientos atípicos.

En conjunto, el análisis exploratorio de compras y ventas declaradas aporta una perspectiva complementaria al estudio de la recaudación tributaria. Mientras la recaudación refleja los ingresos fiscales efectivamente percibidos, las compras y ventas permiten aproximarse a la dinámica económica declarada por los contribuyentes. La lectura conjunta de ambas dimensiones fortalece la construcción de variables analíticas y la interpretación de patrones asociados al riesgo tributario a nivel agregado.

CAPÍTULO 7

MODELADO

El presente capítulo desarrolla la etapa de modelado analítico de la investigación. En esta fase, la información consolidada y depurada en los capítulos anteriores se transforma en una base de perfiles tributarios agregados, construida para aplicar técnicas de aprendizaje no supervisado y detectar patrones diferenciados en el comportamiento de la recaudación.

El propósito del modelado no es identificar contribuyentes individuales ni emitir conclusiones legales sobre perfiles agregados. La información utilizada es pública y agregada, por lo que los resultados deben leerse como señales estadísticas, perfiles de interés o perfiles de prioridad. Esta precisión es importante porque mantiene el enfoque académico del estudio y evita atribuir intenciones o incumplimientos puntuales (*Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future*, s. f.).

Con esta lógica, se ha aplicado técnicas de reducción de dimensionalidad, agrupamiento y detección de anomalías. PCA y UMAP se utilizaron con el fin de comprender y visualizar la estructura de los datos; K-Means, DBSCAN, HDBSCAN y GMM permitieron segmentar perfiles; e Isolation Forest el cual se usó para precisar observaciones atípicas. El capítulo presenta cada etapa con sus resultados y una interpretación orientada al análisis del riesgo tributario.

Para un mayor rigor metodológico, la elección de los modelos se entiende como un conjunto de criterios analíticos exploratorios y no como valores que definen un veredicto. La decisión metodológica busca comparar la utilidad de los modelos e identificar patrones que puedan resultar atípicos. Debido a que el estudio no cuenta con etiquetas institucionales que confirmen evasión fiscal, no es posible validar los resultados mediante métricas supervisadas tradicionales de clasificación. Por esta razón, la selección de parámetros se sustenta en tres

criterios: coherencia estadística, interpretabilidad tributaria y utilidad para diferenciar perfiles agregados dentro de los datos disponibles.

Este criterio resulta especialmente importante en modelos como K-Means, DBSCAN, HDBSCAN, Gaussian Mixture Models e Isolation Forest, ya que cada técnica observa la estructura de los datos desde una lógica distinta. Por tanto, los resultados no se interpretan como evidencia concluyente de evasión fiscal, sino como aproximaciones complementarias para identificar segmentos con comportamiento atípico y prioridad analítica (*AI Principles Overview*, s. f.-a).

7.1 Construcción de la base de modelado

La base de modelado se construyó a partir del conjunto consolidado de recaudación tributaria obtenido de fuentes públicas. Como primera decisión mencionada en el alcance del proyecto se conservaron únicamente los registros correspondientes a sociedades (RUC) y se omitió el año 2026, debido a que su información todavía es parcial. Esta decisión mantiene la homogeneidad temporal del análisis, y enfoca el estudio en el segmento donde existe mayor recaudación como se ha mencionado en capítulos anteriores.

Después, se revisó la variable de actividad económica. En esta etapa se eliminaron las categorías "9" y "NO TIENE", porque no son actividades económicas válidas que se puedan interpretar de mejor forma. Aunque son parte de la base original de data sets, seguirlos utilizando del modelado podía generar ruido y alterar la lectura de los resultados de sectores bien identificados.

Tabla 24
Resumen de construcción de la base de modelado

Etapas	Criterio aplicado	Resultado
Base inicial de recaudación para sociedades	Filtro: sociedades y años 2017-2025	1.604.597 registros y 11 variables
Registros sectoriales excluidos	Códigos de actividad económica "9" y "NO TIENE"	1.625 registros
Perfil analítico final	Agrupación por año, provincia y código de actividad económica	4.377 perfiles y 10 variables base
Matriz de modelado	Selección de variables numéricas y estandarización	4.377 perfiles y 8 variables de entrada

Nota. Elaboración propia a partir del notebook de modelado. La base final corresponde a perfiles agregados por año, provincia y código de actividad económica.

La construcción de perfiles agregados ayuda a pasar de registros operativos individuales de recaudación a unidades que los modelos pueden comparar. Esta transformación tiene como fin dar coherencia al objetivo del trabajo, ya que el análisis se realizó a nivel sectorial y territorial, no a nivel de contribuyentes puntuales.

7.2 Construcción del perfil analítico y variables derivadas

Cada fila del modelo representa un grupo específico: un año, una provincia y un sector económico. Por cada perfil se calcularon medidas de nivel, tendencia central, dispersión y concentración. Esta decisión no sólo muestra cuanto se recauda, sino también cómo se comporta la recaudación dentro las combinaciones de sectores y territorios.

Además de las variables agregadas básicas, se construyeron indicadores derivados como el coeficiente de variación, la participación del valor máximo sobre el total recaudado, el rango de recaudación y se transformó de forma logarítmica el total recaudado. Estos indicadores ayudan ver comportamientos no tan visibles si únicamente nos remitimos al monto total recaudado.

Tabla 25
Variables utilizadas para caracterizar los perfiles tributarios

Variable	Descripción	Uso dentro del modelado
total_recaudado	Suma del valor recaudado por año, provincia y sector.	Representa el peso fiscal total del perfil.
promedio_recaudado	Promedio del valor recaudado en el perfil.	Permite comparar niveles medios de recaudación.
mediana_recaudado	Mediana del valor recaudado.	Reduce el efecto de valores extremos.
std_recaudado	Desviación estándar de la recaudación.	Mide dispersión interna.
cv_recaudacion	Coeficiente de variación de la recaudación.	Mide volatilidad relativa.
peak_share	Participación del valor máximo frente al total recaudado.	Identifica concentración temporal.
rango_recaudado	Diferencia entre el valor máximo y mínimo.	Resume amplitud de variación.
log_total_recaudado	Transformación logarítmica del total recaudado.	Reduce asimetrías de escala.
Registros	Número de registros que conforman el perfil.	Aproxima volumen operativo observado.

Nota. Elaboración propia a partir de las variables construidas en el notebook de modelado. Las variables se calcularon sobre perfiles agregados por año, provincia y sector económico.

La selección de variables combina magnitud, dispersión, concentración y volumen. Estas variables y en esta estructura es la necesaria para un modelo no supervisado porque permite que los algoritmos identifiquen los distintos perfiles no solo por cantidad de recaudación, sino también por volatilidad y estructura de cómo se comportó la recaudación.

7.3 Estandarización de variables

Como paso previo antes de modelar, las variables con las que se va a trabajar fueron estandarizadas mediante *StandardScaler* (*StandardScaler*, s. f.). Esta etapa fue necesaria porque las variables se encuentran en escalas muy diferentes: algunas representan montos monetarios, otras proporciones y otras son cantidades. Sin estandarización, los algoritmos pueden desviarse y dar más peso a variables de mayor magnitud numérica, especialmente aquellas asociadas a montos de recaudación.

La matriz final utilizada para los modelos quedó conformada por 4.377 perfiles y 8 variables numéricas de entrada. Esta matriz de variables es la base común para la reducción de dimensiones, los modelos de agrupamiento y por último la detección de anomalías.

7.4 Análisis de componentes principales (PCA)

El Análisis de Componentes Principales fue la técnica utilizada aplicada para la reducción de dimensionalidad. La finalidad de esto fue resumir la información contenida de todas las variables originales y facilitar la identificación de los perfiles tributarios. En este caso, PCA no se utilizó para construir un puntaje de prioridad analítica, sino para cómo se ordenan los datos y cuánta información importante está resumida en menos dimensiones.

Tabla 26
Varianza explicada por los componentes principales

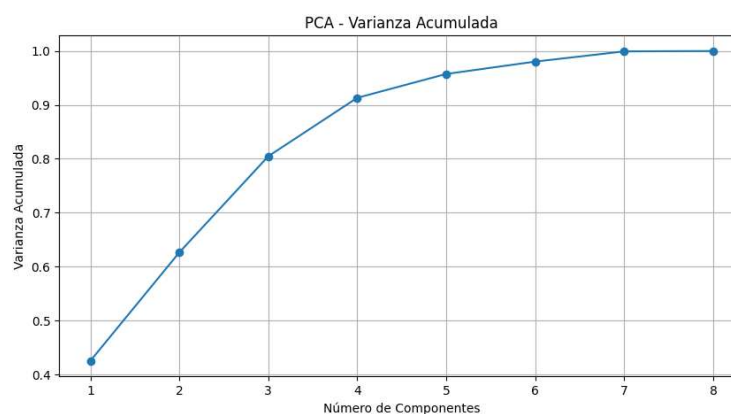
Componente	Varianza	Varianza acumulada
1	42,50 %	42,50 %
2	20,12 %	62,62 %
3	17,85 %	80,47 %
4	10,85 %	91,32 %
5	4,43 %	95,75 %
6	2,29 %	98,04 %

Componente	Varianza	Varianza acumulada
7	1,90 %	99,93 %
8	0,07 %	100,00 %

Nota. Elaboración propia a partir del modelo PCA aplicado sobre las variables estandarizadas. Los valores se expresan como porcentaje de varianza explicada.

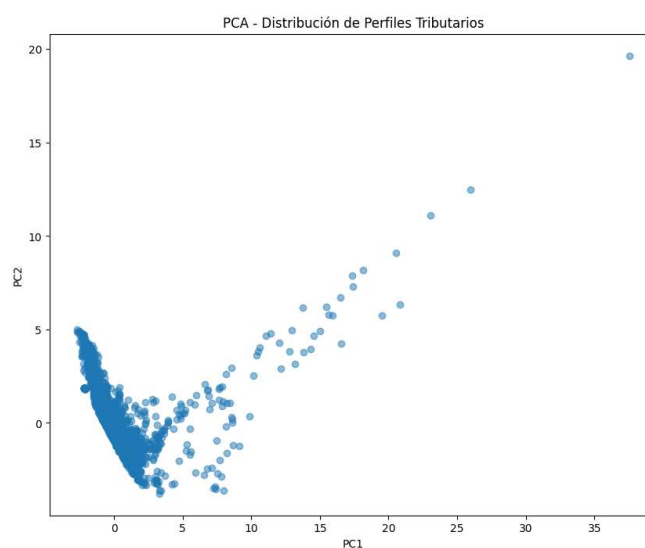
Los tres primeros componentes principales pueden explicar el 80,47 % de la varianza acumulada, mientras que los cuatro primeros logran explicar el 91,32 %. Este resultado indica que una parte importante de la estructura de los datos puede representarse con menos dimensiones o en otras palabras hay pocas variables que tienen peso, y lo hacen sin perder significativamente información relevante del conjunto universo analizado.

Figura 15 Varianza acumulada del modelo PCA



Nota. Elaboración propia a partir del notebook de modelado. La figura muestra la proporción acumulada de varianza explicada al incorporar sucesivos componentes principales

Figura 16 Distribución de perfiles tributarios en el espacio PCA



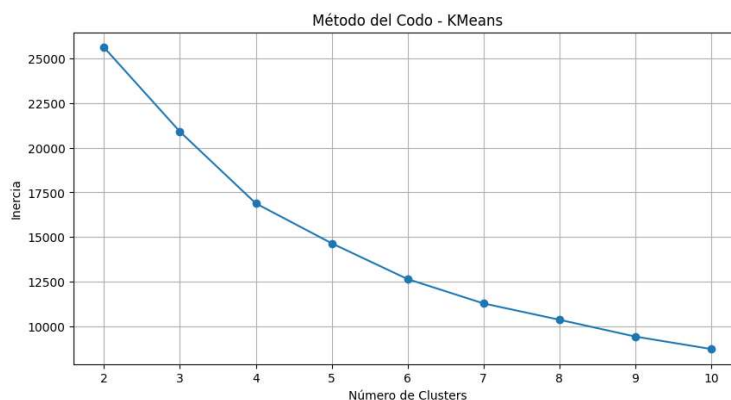
Nota. Elaboración propia a partir del notebook de modelado. Cada punto representa un perfil tributario agregado proyectado en los dos primeros componentes principales.

La gráfica de PCA muestra que existen perfiles concentrados en una zona principal y otros ubicados en regiones cada vez alejadas. Estos últimos no podrían interpretarse automáticamente como evasión, pero si sirve como justificación de la aplicación de modelos de segmentación y detección de anomalías para analizar si corresponden a patrones de tributación diferenciada.

7.5 Segmentación mediante K-Means

K-Means se utilizó para segmentar los grupos con perfiles tributarios de comportamiento similar. Antes de fijar el número de clusters, se revisaron el método del codo y el coeficiente silhouette. El resultado de silhouette muestra que número óptimo $k = 2$ o sea el valor más alto; sin embargo, se trabajó con $k = 5$ porque necesariamente se debe tener mayor granularidad, no se puede segmentar binariamente en el caso de análisis y será más útil desde el punto de vista interpretativo, cuando se tenga los resultados (Winland, 2024).

Esta decisión aparentemente arbitraria se debe entender como una decisión analítica, no como una decisión puramente matemática. Con dos clusters se obtiene una separación binaria no óptima en el presente caso y menos informativa para diferenciar niveles de priorización. Con cinco clusters se logra distinguir perfiles de baja, media, alta y muy alta relevancia analítica, lo cual resulta más útil y coherente en la investigación («(PDF) LQG-Based Virtual Inertial Control of Islanded Microgrid Load Frequency Control and DoS Attack Vulnerability Analysis», 2026).

Figura 17 Método del codo para selección de clusters en K-Means

Nota. Elaboración propia a partir del notebook de modelado. La figura muestra la reducción de inercia al incrementar el número de clusters.

Tabla 27

Coeficiente silhouette para diferentes valores de k

Número de clusters	Silhouette
2	0,8007
3	0,3527
4	0,4198
5	0,4243
6	0,3696
7	0,2580
8	0,2774
9	0,2781
10	0,2843

Nota. Elaboración propia a partir de la evaluación de K-Means sobre la matriz estandarizada. El valor más alto corresponde a $k = 2$, aunque se utiliza $k = 5$ por criterio de interpretabilidad analítica.

Tabla 28

Distribución de perfiles por cluster K-Means

Cluster K-Means	Perfiles
1	3.117
2	825
0	360
4	45
3	30

Nota. Elaboración propia a partir del modelo K-Means con $k = 5$. Los perfiles corresponden a combinaciones agregadas de año, provincia y sector económico.

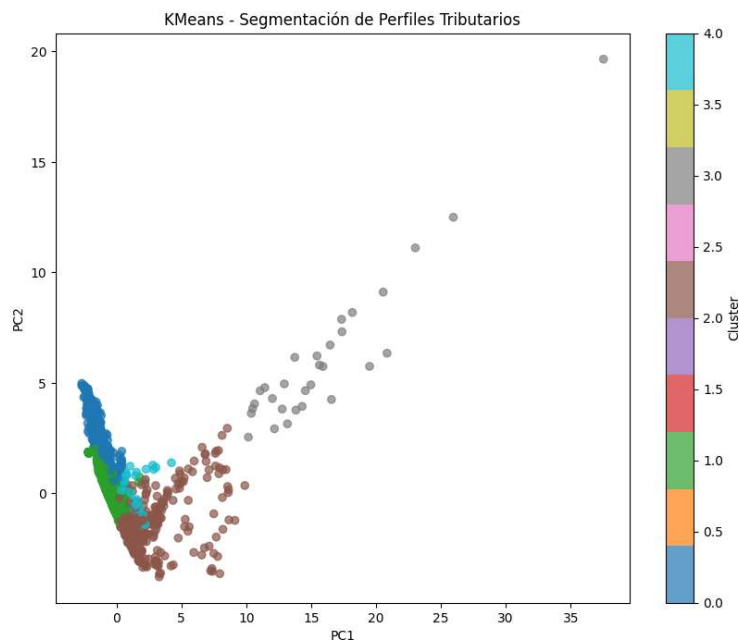
Tabla 29

Perfil promedio de los clusters K-Means

Cluster	Total recaudado promedio (millones USD)	CV promedio	Peak share	Rango promedio (millones USD)	Registros promedio
0	0,24	5,03	0,59	0,13	92,74
1	1,63	2,89	0,11	0,15	257,22
2	100,40	4,89	0,09	6,03	888,52
3	1.467,94	5,00	0,11	108,68	801,53
4	12,05	2,23	0,11	0,82	238,96

Nota. Elaboración propia a partir del notebook de modelado. Los valores monetarios se expresan en millones de dólares para facilitar la lectura.

Figura 18 Segmentación K-Means proyectada en el espacio PCA



Nota. Elaboración propia a partir del notebook de modelado. Los colores representan los clusters identificados por K-Means y proyectados sobre los dos primeros componentes principales.

K-Means nos da lectura inicial de los perfiles tributarios. Los clusters más pequeños, especialmente aquellos con mayores promedios de recaudación y mayor dispersión, son los que tienen mayor interés para un análisis posterior. El modelo nos permite transformar una población grande de perfiles en grupos más entendibles; y esto se logra por la decisión de usar cinco clusters y no como la opción con valor más grande de silhouette.

Aunque el coeficiente silhouette presentó su valor más alto con $k = 2$, esta configuración generaba una separación general del conjunto en 2 perfiles, cercana a una lectura binaria entre un grupo dominante y otro grupo diferenciado. Esta separación puede parecer una solución fuerte desde el punto de vista geométrico, su capacidad explicativa para el análisis tributario es limitada, ya que no permite distinguir niveles intermedios de comportamiento ni construir una priorización analítica gradual.

Por esta razón, se mantuvo la configuración de $k = 5$, cuyo valor de silhouette fue el más alto después de $k = 2$ dentro de los coeficientes alternativos. Esta solución permitió diferenciar perfiles agregados con comportamientos más estables frente a perfiles menos frecuentes, de mayor volumen, dispersión o heterogeneidad. En consecuencia, K-Means no se utilizó para afirmar la existencia de evasión fiscal, sino como una herramienta de segmentación que facilita ordenar los perfiles según similitud estadística y comparar posteriormente sus resultados con modelos basados en densidad, incertidumbre probabilística y detección directa de anomalías.

7.6 Agrupamiento basado en densidad mediante DBSCAN

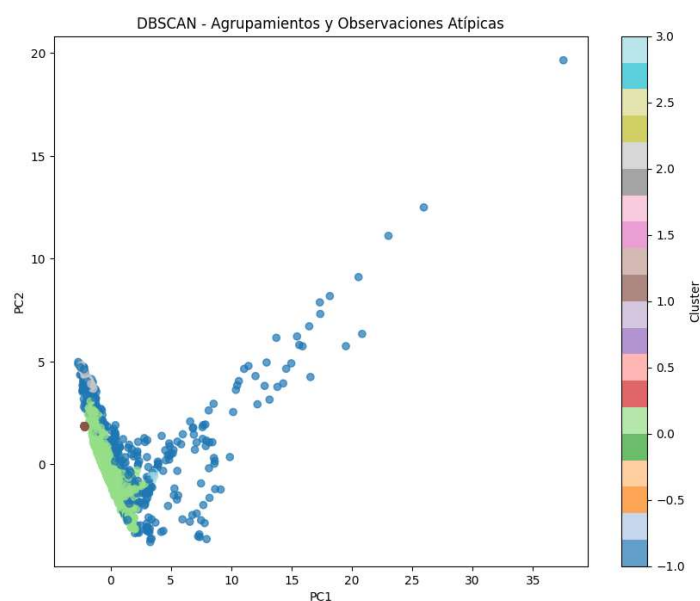
La configuración de DBSCAN se definió con $\text{eps} = 0,7$ y $\text{min_samples} = 10$ como criterio exploratorio con el fin de identificar las zonas más densas dentro de la matriz estandarizada. El parámetro eps es distancia máxima para considerar que dos perfiles son vecinos, mientras que min_samples establece el número mínimo de perfiles cercanos necesarios para formar la densidad de la región. Esta configuración permitió separar la mayoría de los perfiles dentro de agrupamientos principales y, al mismo tiempo, identificar perfiles clasificados como ruido, los cuales no se interpretan como evasión fiscal, sino como segmentos que no siguen la densidad predominante de los datos (Schubert et al., 2017).

En el contexto de este estudio, los perfiles clasificados como ruido por DBSCAN no se interpretan como evidencia de evasión fiscal ni como confirmación de incumplimiento tributario. Su lectura debe ser analítica: representan perfiles agregados que se alejan de las zonas de mayor densidad y que, por ello, pueden ser considerados segmentos con mayor foco de atención para estudios posteriores. Esta interpretación permite utilizar DBSCAN como una técnica complementaria a K-Means, aportando una lectura basada en densidad y no únicamente en cercanía a centroides.

Tabla 30*Distribución de perfiles por cluster DBSCAN*

Cluster DBSCAN	Perfiles
0	3.888
-1	386
1	61
2	34
3	8

Nota. Elaboración propia a partir del modelo DBSCAN aplicado sobre la matriz estandarizada. El cluster -1 representa perfiles clasificados como ruido.

Figura 19 Agrupamiento DBSCAN proyectado en el espacio PCA

Nota. Elaboración propia a partir del notebook de modelado. La figura muestra los clusters y los perfiles clasificados como ruido por el algoritmo.

DBSCAN ha puesto 386 perfiles como ruido. Este resultado es relevante porque estos perfiles no tienen la densidad predominante que poseen los otros datos. Sin embargo, cabe aclarar que esto no equivale a evasión fiscal; únicamente que estos por su configuración de variables, merecen una revisión más puntual y analítica.

7.7 Agrupamiento jerárquico por densidad mediante HDBSCAN

La configuración de HDBSCAN se definió con `min_cluster_size = 30` y `min_samples = 10`. El primer parámetro es para que el tamaño mínimo razonable para considerar la formación de un agrupamiento estable, evitando que grupos demasiado pequeños fueran

interpretados como estructuras relevantes por sí mismos. El segundo parámetro conserva un criterio de densidad suficiente para identificar perfiles que no se integraban de forma consistente a los agrupamientos principales.

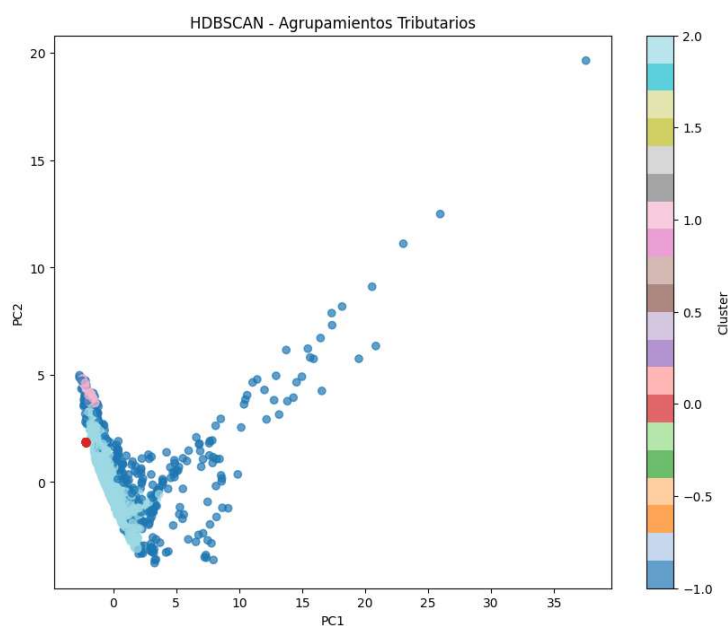
La aplicación de HDBSCAN es necesario por la necesidad de complementar la lectura de DBSCAN y su capacidad de poder reconocer agrupamientos con densidades variables. Esta característica se aplica en los datos organizados por año, provincia y sector económico, ya que no todos los territorios ni actividades económicas presentan el mismo nivel de concentración o volumen de registros. Cabe aclarar que, los perfiles clasificados como ruido por HDBSCAN no se pueden interpretar como evidencia de evasión fiscal, sino como segmentos agregados para un estudio más analítico y profundo (*The hdbscan Clustering Library — hdbscan 0.8.1 documentation*, s. f.).

Tabla 31
Distribución de perfiles por cluster HDBSCAN

Cluster HDBSCAN	Perfiles
2	3.930
-1	344
0	61
1	42

Nota. Elaboración propia a partir del modelo HDBSCAN aplicado sobre la matriz estandarizada. El cluster -1 representa perfiles clasificados como ruido.

Figura 20 Agrupamiento HDBSCAN proyectado en el espacio PCA



Nota. Elaboración propia a partir del notebook de modelado. La figura muestra la separación de perfiles según los clusters obtenidos por HDBSCAN.

HDBSCAN identificó 344 perfiles como ruido, una cantidad ligeramente menor a la obtenida por DBSCAN. Esta casi coincidencia entre modelos basados en densidad da fuerza a la lectura de que hay un subconjunto de perfiles con comportamiento diferente. La interpretación al igual que en los modelos anteriores debe mantenerse como señal, no como prueba de incumplimiento.

7.8 Modelado probabilístico mediante Gaussian Mixture Models

Gaussian Mixture Models se utilizó para asignar perfiles a clusters desde una perspectiva probabilística. A diferencia de K-Means, este modelo no solo asigna una observación a un grupo, sino que también estima la probabilidad de pertenencia a cada componente. Esta característica permite identificar perfiles cuya pertenencia a un único grupo no es clara, es decir, observaciones ubicadas en zonas intermedias entre varios segmentos.

La agregación de Gaussian Mixture Models se utiliza como complemento probabilístico a la segmentación obtenida mediante K-Means. Mientras K-Means asigna cada perfil agregado a un único cluster de forma directa, GMM permite estimar el nivel de pertenencia de cada observación a los distintos componentes del modelo. Esta característica resulta útil para identificar perfiles cuya asignación de segmento no es necesariamente clara y que, en consecuencia, se ubican en zonas de límites de segmentos y distintos comportamientos tributarios agregados (Noble, 2025b).

En este estudio se utilizó la configuración de cinco componentes, para tener una coherencia comparativa con la segmentación aplicada K-Means. De igual forma como los modelos anteriores, los componentes de GMM no se interpretan como categorías de clasificación ni como grupos de evasión fiscal, sino que la estructura común probabilística de cada grupo permite observar la incertidumbre de pertenencia de los perfiles analizados. Es por

eso por lo que, GMM aporta una forma de lectura que complementa a K-Means y su orientación es identificar perfiles con menor claridad de asignación dentro del espacio de variables modeladas.

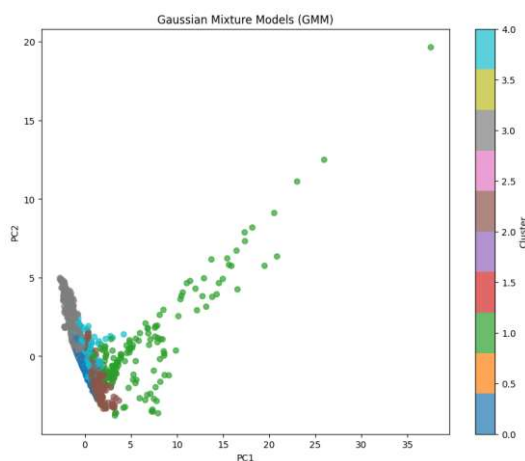
Tabla 32

Distribución de perfiles por cluster GMM

Cluster GMM	Perfiles
0	1.611
3	1.327
4	776
2	505
1	158

Nota. Elaboración propia a partir del modelo Gaussian Mixture Models con cinco componentes.

Figura 21 Clusters GMM proyectados en el espacio PCA



Nota. Elaboración propia a partir del notebook de modelado. La figura muestra la asignación probabilística de clusters representada sobre los componentes principales.

GMM hace su aporte desde la probabilidad a diferencia de los modelos anteriores porque permite reducir la incertidumbre de pertenencia. Los perfiles con baja probabilidad no necesariamente son extremos en monto, pero sí pueden tener una estructura entre lo ambiguo y lo explicable. Definitivamente esto es útil para el análisis tributario porque algunos perfiles pueden no ser identificados como normales, ni necesariamente como atípicos, sino que están en el borde de ambas etiquetas.

7.9 Detección directa de anomalías mediante Isolation Forest

Se aplicó Isolation Forest ya como modelo específico de detección de anomalías. Su lógica es la de diferenciar observaciones mediante particiones aleatorias: los perfiles más atípicos por lo general requieren menos cortes para quedar aislados del resto. Se utilizó una contaminación de 0,08 como decisión metodológica como un umbral exploratorio no como una verdad inamovible, con esto tratamos de lograr una proporción controlada de perfiles anómalos.

La parametrización de Isolation Forest se estableció con $\text{contamination} = 0,08$ como un umbral exploratorio coherente para priorizar una proporción reducida de perfiles agregados con comportamiento atípico, sin que el criterio sea demasiado restrictivo ni demasiado amplio. Este valor no representa una tasa real de evasión fiscal ni una estimación de incumplimiento tributario, sino un umbral metodológico utilizado para identificar perfiles que se aíslan con mayor facilidad frente a los comportamientos más regulares (*IsolationForest*, s. f.).

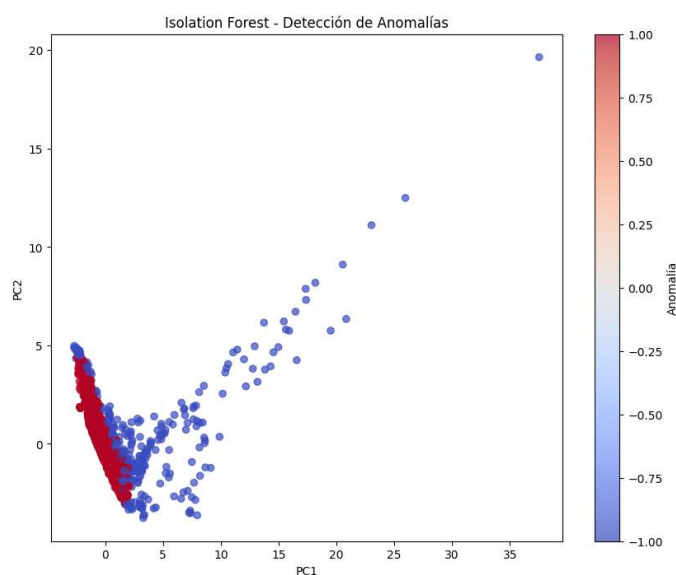
En este contexto, los perfiles señalados como anómalos por Isolation Forest deben interpretarse como segmentos de alta prioridad analítica y no como confirmación de evasión fiscal. Su utilidad dentro del estudio consiste en aportar una lectura directa de atipicidad, complementaria a la segmentación por similitud de K-Means, la densidad de DBSCAN y HDBSCAN, y la incertidumbre de pertenencia observada mediante GMM.

Tabla 33
Distribución de perfiles normales y anómalos según Isolation Forest

Clasificación Isolation Forest	Perfiles
Normal (1)	4.026
Anómalo (-1)	351

Nota. Elaboración propia a partir del modelo Isolation Forest aplicado sobre la matriz estandarizada. La

categoría -1 corresponde a perfiles clasificados como anómalos por el algoritmo.

Figura 22 Anomalías detectadas por Isolation Forest en el espacio PCA

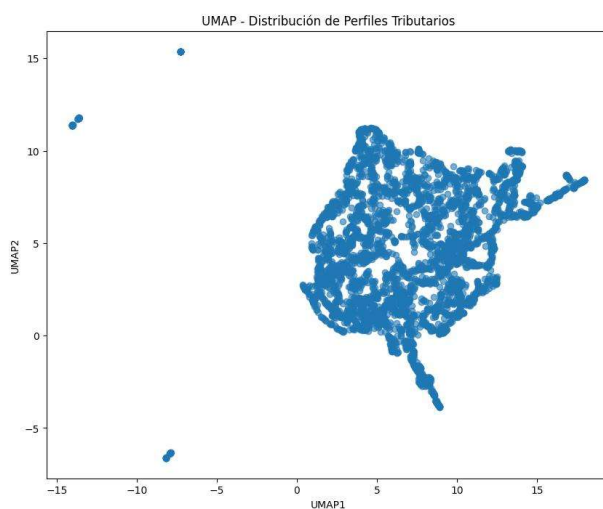
Nota. Elaboración propia a partir del notebook de modelado. La figura muestra los perfiles normales y anómalos proyectados sobre los dos primeros componentes principales.

Isolation Forest identificó 351 perfiles anómalos y mostró coherencia con los modelos de agrupamiento aplicados previamente. Este resultado es útil porque el modelo está orientado a identificar directamente perfiles atípicos dentro del espacio de variables. Aun así, esta rareza estadística no debe interpretarse como evasión fiscal: se trata de una señal para priorizar revisiones analíticas, en concordancia con la lógica de los otros modelos.

7.10 Visualización no lineal mediante UMAP

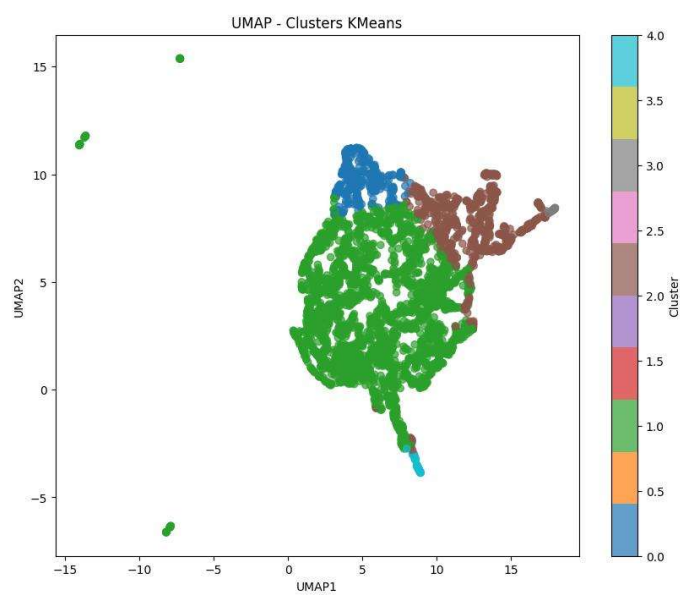
UMAP se utilizó como técnica de reducción de dimensionalidad no lineal. Con el fin de mostrar visualmente relaciones que resulta más complejas entre perfiles que los que puede capturarse con PCA. En este trabajo, UMAP no genera por sí mismo un puntaje ni una clasificación de riesgo; su utilidad se concentra en validar visualmente la separación de clusters y la ubicación de anomalías detectadas por otros modelos.

Figura 23 Distribución general de perfiles tributarios mediante UMAP



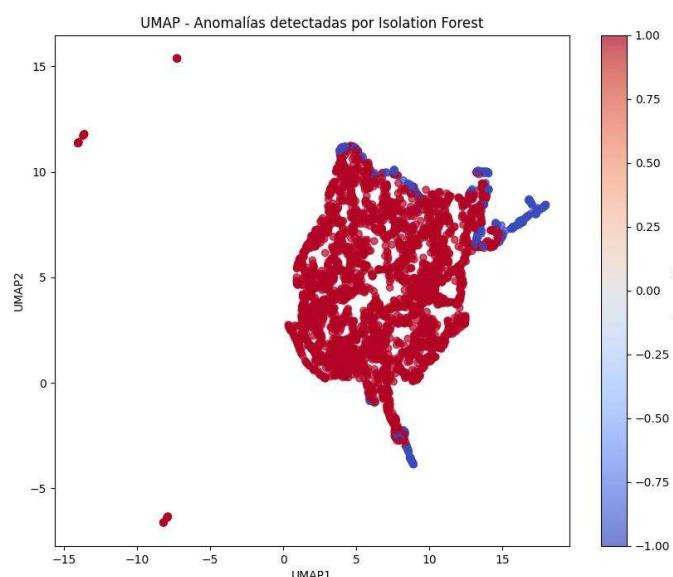
Nota. Elaboración propia a partir del notebook de modelado. La figura representa los perfiles tributarios en un espacio bidimensional no lineal.

Figura 24 Clusters K-Means proyectados en el espacio UMAP



Nota. Elaboración propia a partir del notebook de modelado. La figura permite observar la separación de los clusters K-Means en el espacio UMAP.

Figura 25 Anomalías de Isolation Forest proyectadas en el espacio UMAP



Nota. Elaboración propia a partir del notebook de modelado. La figura muestra la ubicación de los perfiles anómalos identificados por Isolation Forest dentro del espacio UMAP.

UMAP permite revisar de forma visual si los perfiles priorizados se ubican en zonas diferenciadas del espacio de datos y resaltados con colores, entre más tono de azul lo más anómalo. Esta visualización complementa los resultados numéricos y ayuda a explicar el comportamiento de los modelos, pero no sustituye la interpretación estadística ni la validación ya con variables tributarias.

7.11 Análisis de sensibilidad metodológica

Como parte del proceso de modelado, los parámetros utilizados en los algoritmos no supervisados se interpretaron como decisiones metodológicas exploratorias y no como valores óptimos definitivos. Esta consideración es necesaria debido a que la base de datos utilizada no cuenta con etiquetas institucionales que permitan confirmar evasión fiscal, por lo que no es posible validar los modelos mediante métricas supervisadas tradicionales (*¿Qué es el aprendizaje no supervisado? Google Cloud | Google Cloud, s. f.*).

En este sentido, la elección de $k = 5$ en K-Means, $\text{eps} = 0,7$ y $\text{min_samples} = 10$ en DBSCAN, $\text{min_cluster_size} = 30$ y $\text{min_samples} = 10$ en HDBSCAN, cinco componentes en

GMM y contamination = 0,08 en Isolation Forest respondió a un criterio analítico empírico por observación de los datos que poseíamos y lograr comparación entre modelos y con utilidad analítica. Cada parámetro permitió observar los perfiles agregados desde una lógica distinta: similitud estadística, densidad, estabilidad de agrupamientos, incertidumbre de pertenencia y atipicidad directa.

Por tanto, los resultados obtenidos no deben entenderse como una clasificación concluyente de evasión fiscal, sino como señales complementarias para identificar segmentos agregados que presentan comportamientos diferenciados respecto del patrón dominante. Esta lectura permite fortalecer la utilidad del modelo como herramienta de priorización macroanalítica y deja abierta la posibilidad de validación futura con datos internos, trazabilidad institucional y registros confirmados por la administración tributaria.

7.12 Síntesis del capítulo

El proceso de modelado ha permitido transformar la base de recaudación en una matriz de perfiles tributarios, y mediante técnicas no de aprendizaje no supervisado poder observar su estructura interna. En los resultados se aprecia grupos claramente agrupados, perfiles con mayor volumen o dispersión y perfiles que algunos modelos los van clasificando como ruido o anomalía.

La combinación de modelos fue útil porque cada técnica mira el problema desde un ángulo distinto. K-Means ordena la población en segmentos interpretables; DBSCAN y HDBSCAN identifican perfiles que no siguen la densidad dominante; GMM revela incertidumbre de grupo; e Isolation Forest detecta anomalías de manera ya más directa. Entre todos estos, nos proporcionan la base sólida para construir los puntajes o puntuaciones, señales y criterios de priorización analítica en el capítulo siguiente (2.7. *Novelty and Outlier Detection*, s. f.).

CAPÍTULO 8

PUNTAJE Y PRIORIZACIÓN ANALÍTICA DEL RIESGO

El presente capítulo desarrolla la etapa de puntuación y priorización analítica de los perfiles tributarios agregados. Una vez aplicados los modelos no supervisados en el capítulo anterior, sus resultados se convirtieron en indicadores interpretativos que permitieron ordenar los perfiles y a su vez la lectura de estos, según su nivel de interés analítico. Esta traducción se realizó mediante puntajes de prioridad, señales tributarias y categorías de priorización.

Es necesario aclarar que el puntaje construido en este capítulo no representa una probabilidad de evasión fiscal ni una clasificación definitiva de incumplimiento tributario. Su objetivo es ordenar perfiles agregados según patrones detectados por los modelos y señales asociadas a recaudación, volatilidad, dispersión, concentración temporal y volumen de registros. De esta manera, los resultados deben interpretarse como una herramienta de priorización macroanalítica y no como evidencia concluyente sobre contribuyentes individuales (*AI Principles Overview*, s. f.-b).

8.1 Puntaje de prioridad analítica basado en K-Means

Para K-Means se construyó un puntaje de prioridad analítica a partir de la asignación ordinal de valores a los clusters obtenidos. Esta asignación permitió traducir la segmentación estadística en una escala interpretativa de prioridad, sin modificar el entrenamiento original del modelo ni incorporar etiquetas externas de evasión fiscal. Los clusters con menor relevancia analítica recibieron un valor de 25, el cluster intermedio recibió un valor de 50, el cluster de mayor relevancia relativa recibió un valor de 75 y el cluster con comportamiento más heterogéneo recibió un valor de 100.

Esta escala no representa una probabilidad real de evasión ni una clasificación legal de riesgo tributario. Su función es facilitar la lectura comparativa de los perfiles agregados,

permitiendo ordenar los grupos según su comportamiento relativo dentro de las variables analizadas. En este sentido, el puntaje de prioridad analítica basado en K-Means constituye una capa posterior de interpretación y no una prueba concluyente de incumplimiento.

Tabla 34*Distribución de perfiles según puntaje K-Means*

Puntaje de prioridad K-Means	Perfiles
25	3.477
75	825
50	45
100	30

Nota. Elaboración propia a partir de la asignación de puntaje de prioridad analítica sobre los clusters K-Means.

Tabla 35*Distribución de perfiles según nivel K-Means*

Nivel	Perfiles
Bajo	3.477
Alto	825
Medio	45
Muy Alto	30

Nota. Elaboración propia a partir de la clasificación del puntaje de prioridad K-Means en niveles bajo, medio, alto y muy alto.

La concentración de la mayoría de los perfiles en el nivel bajo es esperable dentro de un enfoque orientado a identificar comportamientos anómalos. Los perfiles clasificados como alto o muy alto no representan de ninguna manera evasión tributaria; representan segmentos que, por su posición relativa dentro del espacio de variables, merecen mayor revisión analítica. Esta lectura evita transformar el puntaje se pueda interpretar como un juicio, sino más bien como un nivel para priorizar análisis.

8.2 Señales de riesgo tributario

Adicionalmente al puntaje de prioridad analítica basado en K-Means, se construyeron señales tributarias basadas en umbrales estadísticos. Para ello se utilizó el percentil 90 de cada variable, lo que significa que una señal se activa únicamente cuando el perfil se ubica dentro del 10 % superior de la distribución. Las señales consideradas fueron: alta recaudación, alta volatilidad, concentración temporal, alta dispersión y alto volumen de registros.

Estas señales no formaron parte del entrenamiento de los modelos, sino que se incorporaron como una capa posterior de interpretación. Su intención es la de identificar perfiles que presentan valores especialmente altos en variables relevantes para el análisis tributario. Por ello, la acumulación de señales debe entenderse que son decisiones para poder a posterior enfocar un interés analítico y no como una prueba de evasión fiscal.

Tabla 36*Resumen de señales de riesgo tributario identificadas*

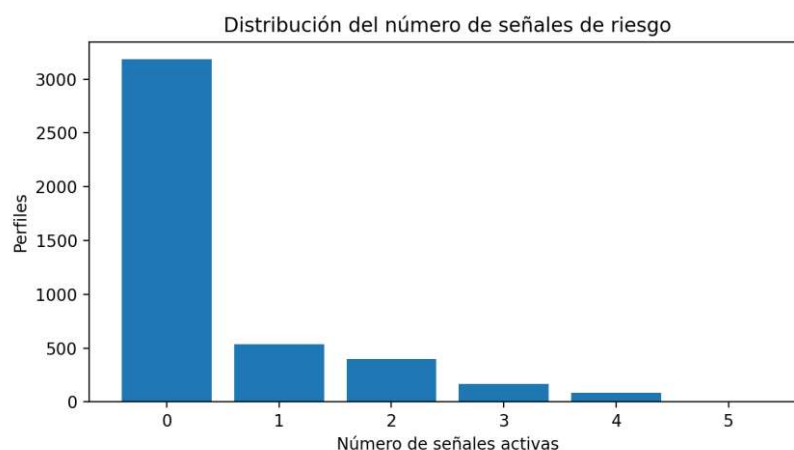
Señal	Perfiles
Alta recaudación	438
Alta volatilidad	438
Concentración temporal	438
Alta dispersión	438
Alto volumen de registros	438

Nota. Elaboración propia a partir de umbrales definidos en el percentil 90 de las variables de modelado. Cada señal identifica perfiles que superan el comportamiento general de la población analizada.

Tabla 37*Distribución del número total de señales por perfil*

Número de señales activas	Perfiles
0	3.187
1	537
2	397
3	167
4	87
5	2

Nota. Elaboración propia a partir de la suma de señales activas por perfil tributario.

Figura 26 Distribución del número de señales de riesgo por perfil

Nota. Elaboración propia a partir de los resultados del notebook de modelado. La figura muestra cuántos perfiles acumulan cero, una o más señales de riesgo.

La mayor parte de los perfiles no presenta señales tributarias acumuladas, lo cual confirma que el criterio del percentil 90 funciona como un umbral para selección. Sin embargo, existe un grupo reducido que acumula tres o más señales simultáneamente. Estos perfiles no deben interpretarse como evidencia de evasión fiscal, sino como segmentos agregados que combinan varias condiciones poco frecuentes dentro del universo observado y que, por tanto, serán más prioritarios para un análisis posterior.

8.3 Priorización analítica con K-Means

La priorización K-Means La priorización analítica basada en K-Means combinó el nivel de prioridad asignado a cada cluster con el número total de señales tributarias activadas por cada perfil. Los perfiles fueron clasificados con prioridad alta cuando pertenecían al nivel muy alto o cuando acumulaban tres o más señales. Se clasificaron con prioridad media aquellos perfiles ubicados en nivel alto o con dos señales activadas, mientras que los perfiles restantes fueron clasificados con prioridad baja.

Esta regla permitió integrar dos lecturas complementarias: por un lado, la segmentación estadística obtenida mediante K-Means; por otro, la acumulación de señales tributarias ubicadas en el 10 % superior de variables relevantes. Por tanto, la prioridad alta no representa evasión confirmada, sino una categoría de ordenamiento analítico para identificar perfiles agregados que requieren mayor revisión.

Tabla 38
Distribución de prioridad analítica según K-Means

Prioridad	Perfiles
Baja	3.361
Media	746
Alta	270

Nota. Elaboración propia a partir de la combinación entre nivel K-Means y número de señales de riesgo activas.

Tabla 39
Principales perfiles priorizados según K-Means

Año	Provincia	Sector	Nivel	Señales	Prioridad	Total recaudado (millones USD)
2018	PICHINCHA	H	Muy Alto	5	Alta	346,42
2020	GUAYAS	N	Alto	5	Alta	55,27
2025	PICHINCHA	G	Muy Alto	4	Alta	3.926,96
2024	PICHINCHA	G	Muy Alto	4	Alta	3.585,32
2023	PICHINCHA	G	Muy Alto	4	Alta	2.968,08
2019	PICHINCHA	G	Muy Alto	4	Alta	2.068,39
2017	PICHINCHA	G	Muy Alto	4	Alta	1.906,74
2020	PICHINCHA	G	Muy Alto	4	Alta	1.670,38
2022	GUAYAS	G	Alto	4	Alta	1.475,84
2021	GUAYAS	G	Alto	4	Alta	1.284,75

Nota. Elaboración propia a partir de los perfiles con mayor acumulación de señales y mayor recaudación total.

Los valores monetarios se expresan en millones de dólares.

K-Means priorizó 270 perfiles para que se ponga atención prioritaria o alta. Entre los principales perfiles aparecen perfiles de Pichincha y Guayas, especialmente en sectores como comercio, manufactura, transporte y actividades financieras. Esta concentración era de esperarse, porque estas provincias y sectores concentran la mayor participación económica; por ello, hay que leer bien esta parte, de hecho, no implica irregularidad real en sí, sino entender que allí se concentran perfiles con mayor peso y mayor diferenciación estadística comparada con el resto de las provincias.

8.4 Priorización basada en modelos de densidad

En DBSCAN y HDBSCAN el nivel de priorización fue elaborado a partir de la clasificación de ruido que estos poseían. Los perfiles asignados al cluster -1 recibieron una puntuación de 100 y una prioridad alta, mientras que el resto recibió una puntuación de 0 y prioridad baja. Este criterio es de carácter binario, por lo que su interpretación no debe darse como cumple o no cumple bien las obligaciones tributarias por ende hay q ser cauteloso con afirmaciones.

Tabla 40*Puntaje, nivel y prioridad según DBSCAN*

Dimensión	Categoría	Perfiles
Puntaje DBSCAN	0	3.991
Puntaje DBSCAN	100	386
Nivel DBSCAN	Bajo	3.991
Nivel DBSCAN	Alto	386
Prioridad DBSCAN	Baja	3.991
Prioridad DBSCAN	Alta	386

Nota. Elaboración propia a partir de los perfiles clasificados como ruido por DBSCAN.

Tabla 41*Principales perfiles priorizados según DBSCAN*

Año	Provincia	Sector	Puntaje	Prioridad	Total recaudado (millones USD)
2025	PICHINCHA	G	100	Alta	3.926,96
2024	PICHINCHA	G	100	Alta	3.585,32
2023	PICHINCHA	G	100	Alta	2.968,08
2025	PICHINCHA	K	100	Alta	2.079,44
2018	PICHINCHA	G	100	Alta	2.078,09
2019	PICHINCHA	G	100	Alta	2.068,39
2024	PICHINCHA	K	100	Alta	2.065,60
2025	GUAYAS	G	100	Alta	1.930,84
2017	PICHINCHA	G	100	Alta	1.906,74
2022	PICHINCHA	G	100	Alta	1.906,02

Nota. Elaboración propia a partir de los perfiles clasificados como ruido por DBSCAN y ordenados por

recaudación total. Los valores monetarios se expresan en millones de dólares.

Tabla 42*Puntaje, nivel y prioridad según HDBSCAN*

Dimensión	Categoría	Perfiles
Puntaje HDBSCAN	0	4.033
Puntaje HDBSCAN	100	344
Nivel HDBSCAN	Bajo	4.033
Nivel HDBSCAN	Alto	344
Prioridad HDBSCAN	Baja	4.033
Prioridad HDBSCAN	Alta	344

Nota. Elaboración propia a partir de los perfiles clasificados como ruido por HDBSCAN.

Tabla 43*Principales perfiles priorizados según HDBSCAN*

Año	Provincia	Sector	Puntaje	Prioridad	Total recaudado (millones USD)
2025	PICHINCHA	G	100	Alta	3.926,96
2024	PICHINCHA	G	100	Alta	3.585,32
2023	PICHINCHA	G	100	Alta	2.968,08
2025	PICHINCHA	K	100	Alta	2.079,44
2018	PICHINCHA	G	100	Alta	2.078,09
2019	PICHINCHA	G	100	Alta	2.068,39
2024	PICHINCHA	K	100	Alta	2.065,60
2025	GUAYAS	G	100	Alta	1.930,84
2017	PICHINCHA	G	100	Alta	1.906,74
2022	PICHINCHA	G	100	Alta	1.906,02

Nota. Elaboración propia a partir de los perfiles clasificados como ruido por HDBSCAN y ordenados por

recaudación total. Los valores monetarios se expresan en millones de dólares.

DBSCAN identificó 386 perfiles de alta prioridad y HDBSCAN identificó 344. La cercanía entre ambos resultados sugiere que una parte relevante de los perfiles priorizados presenta baja integración a las zonas de mayor densidad del espacio de variables. En otras palabras, estos perfiles no se agrupan de manera clara con el comportamiento predominante observado en la base de modelado.

Esta coincidencia no es realmente una evidencia de evasión fiscal, sino como una señal analítica de comportamientos diferentes. La utilidad de los modelos basados en densidad consiste en identificar perfiles agregados que se ubican fuera de las estructuras más frecuentes, lo cual permite complementar la priorización obtenida mediante K-Means y contrastarla con técnicas probabilísticas y de detección directa de anomalías.

8.5 Priorización basada en incertidumbre probabilística con GMM

En el caso de GMM, el puntaje se construyó a partir de la incertidumbre de pertenencia de cada perfil a un componente del modelo. Cuando un perfil presenta menor certeza de asignación a un único grupo, se interpreta que sus características se ubican en una zona intermedia entre varios segmentos. Esta condición no representa evasión fiscal ni anomalía confirmada, sino una señal de ambigüedad probabilística útil para el análisis comparativo.

Tabla 44
Estadísticos descriptivos del puntaje GMM

Estadístico	Puntaje GMM
Count	4.377,00
Mean	2,5034
Std	7,9259
Min	0,0000
25%	0,0000
50%	0,0012
75%	0,1988
Max	49,5648

Nota. Elaboración propia a partir del puntaje definido como uno menos la probabilidad máxima de pertenencia al cluster, multiplicado por 100.

Tabla 45*Señales y prioridad según GMM*

Dimensión	Categoría	Perfiles
Señal GMM	0	3.910
Señal GMM	1	467
Prioridad GMM	Baja	3.910
Prioridad GMM	Alta	467

Nota. Elaboración propia a partir de la clasificación de perfiles con probabilidad máxima de pertenencia inferior a 0,95.

Tabla 46*Principales perfiles priorizados según GMM*

Año	Provincia	Sector	Probabilidad GMM	Puntaje GMM	Prioridad
2024	EL ORO	A	0,504	49,56	Alta
2021	SANTA ELENA	C	0,506	49,39	Alta
2017	ORELLANA	J	0,509	49,09	Alta
2022	AZUAY	J	0,510	49,02	Alta
2024	COTOPAXI	N	0,510	48,98	Alta
2024	BOLIVAR	G	0,513	48,69	Alta
2025	MANABI	L	0,513	48,65	Alta
2019	NAPO	A	0,514	48,61	Alta
2024	PASTAZA	H	0,515	48,48	Alta
2024	MANABI	H	0,515	48,47	Alta

Nota. Elaboración propia a partir de los perfiles con menor probabilidad máxima de pertenencia a un único cluster.

GMM priorizó 467 perfiles con menor certeza de pertenencia a un único componente. Este resultado significa que GMM responde a una lógica distinta: identificar perfiles cuya asignación probabilística no es completamente clara. Así que podemos decir que, estos perfiles deben interpretarse como segmentos de clasificación incierta, no necesariamente como perfiles extremos o disonantes en términos de recaudación.

Esta lectura complementa los resultados de los modelos anteriores. Mientras K-Means permite segmentar por similitud estadística, DBSCAN y HDBSCAN identifican baja integración a zonas densas, y GMM permite observar perfiles ubicados entre varios comportamientos posibles. En consecuencia, la prioridad generada por GMM debe entenderse como una señal analítica de incertidumbre y no como evidencia de incumplimiento tributario.

8.6 Priorización mediante Isolation Forest

Para Para Isolation Forest, los perfiles clasificados como anómalos recibieron un puntaje de prioridad de 100, nivel alto y prioridad alta. Los perfiles clasificados como regulares recibieron un puntaje de 0, nivel bajo y prioridad baja. Esta regla mantiene una estructura binaria porque responde a la lógica propia del algoritmo, diseñado para identificar observaciones que se aíslan con mayor facilidad respecto del comportamiento predominante del conjunto de datos.

Esta clasificación no debe entenderse como una confirmación de evasión fiscal. El resultado indica únicamente que ciertos perfiles agregados presentan mayor atipicidad estadística frente al universo analizado y por tanto, pueden ser considerados segmentos de interés para una revisión posterior.

Tabla 47

Puntaje, nivel y prioridad según Isolation Forest

Dimensión	Categoría	Perfiles
Puntaje Isolation Forest	0	4.026
Puntaje Isolation Forest	100	351
Nivel Isolation Forest	Bajo	4.026
Nivel Isolation Forest	Alto	351
Prioridad Isolation Forest	Baja	4.026
Prioridad Isolation Forest	Alta	351

Nota. Elaboración propia a partir de la clasificación normal/anómala generada por Isolation Forest.

Tabla 48

Principales perfiles priorizados según Isolation Forest

Año	Provincia	Sector	Puntaje IF	Puntaje de priorización	Prioridad	Total recaudado (millones USD)
2018	PICHINCHA	B	-0,288	100	Alta	1.191,97
2023	PICHINCHA	B	-0,285	100	Alta	854,98
2019	PICHINCHA	K	-0,285	100	Alta	1.537,02
2020	PICHINCHA	K	-0,280	100	Alta	1.351,15
2018	PICHINCHA	K	-0,277	100	Alta	1.453,85
2022	PICHINCHA	K	-0,276	100	Alta	1.715,69
2024	PICHINCHA	K	-0,276	100	Alta	2.065,60
2024	PICHINCHA	O	-0,274	100	Alta	745,58
2021	PICHINCHA	K	-0,274	100	Alta	1.536,34
2022	PICHINCHA	B	-0,272	100	Alta	1.765,79

Nota. Elaboración propia a partir de los perfiles anómalos ordenados por el puntaje interno del modelo. Los

valores monetarios se expresan en millones de dólares.

Isolation Forest priorizó 351 perfiles agregados. A diferencia de GMM, que identifica perfiles con menor certeza de pertenencia a un único componente, Isolation Forest prioriza observaciones que se aíslan con mayor facilidad dentro del conjunto de datos. Por ello, su resultado aporta una lectura directa de los datos atípicos.

La utilidad de este modelo consiste en complementar los enfoques anteriores mediante una técnica específicamente orientada a la detección de anomalías. Sin embargo, los perfiles priorizados no deben interpretarse como confirmación de evasión fiscal, sino como segmentos agregados que presentan comportamientos suficientemente diferenciados para justificar una mayor atención analítica.

8.7 Comparación general de priorización por modelo

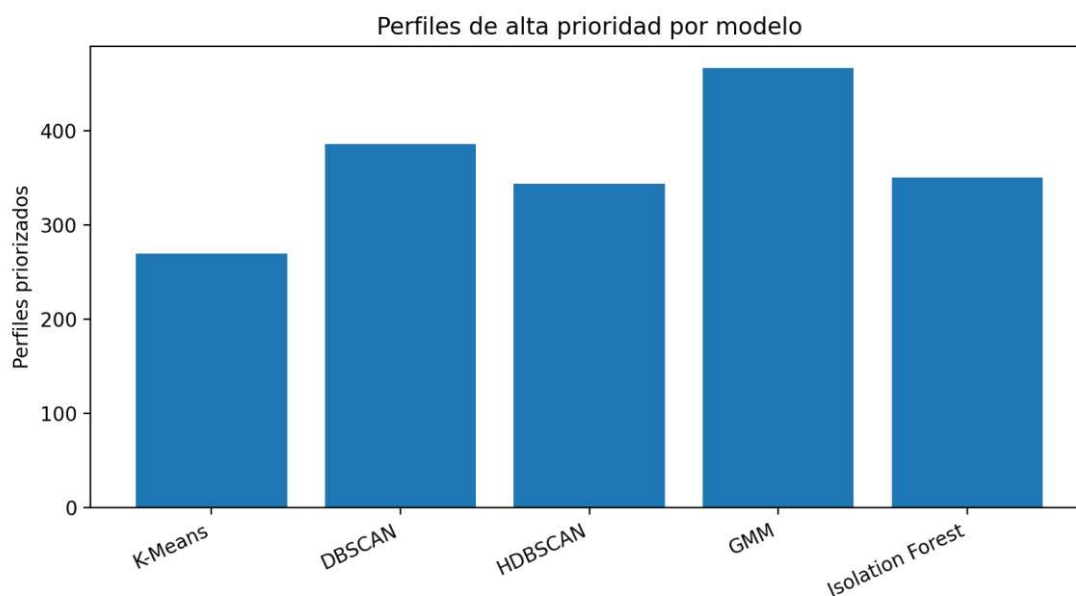
La comparación entre modelos no debe interpretarse como una competencia para escoger un único algoritmo ganador, sino como una lectura complementaria de distintas dimensiones del comportamiento tributario agregado. K-Means aporta interpretabilidad mediante perfiles promedio; DBSCAN y HDBSCAN identifican observaciones fuera de zonas densas; GMM incorpora incertidumbre de pertenencia; e Isolation Forest prioriza perfiles que se aíslan con mayor facilidad en el espacio de variables.

Por tanto, la coincidencia entre modelos fortalece la prioridad analítica de un perfil, mientras que las diferencias entre modelos ayudan a comprender la naturaleza específica del comportamiento atípico observado. En este sentido, el valor del enfoque no está en que todos los modelos produzcan exactamente la misma lista de perfiles, sino en que sus resultados permitan contrastar similitud, densidad, incertidumbre y atipicidad desde perspectivas complementarias.

Tabla 49*Comparación de criterios de priorización por modelo*

Modelo	Criterio de priorización	Perfiles de alta prioridad	Lectura analítica
K-Means	Cluster K-Means + señales por percentil 90	270	Prioriza perfiles de alta recaudación, alta dispersión y acumulación de señales.
DBSCAN	Ruido o puntos sin densidad suficiente	386	Identifica observaciones que no se integran a los grupos principales.
HDBSCAN	Ruido en agrupamiento jerárquico por densidad	344	Similar a DBSCAN, pero más flexible ante densidades diferentes.
GMM	Incertidumbre de pertenencia al cluster	467	Detecta perfiles de clasificación ambigua entre segmentos.
Isolation Forest	Aislamiento directo de anomalías	351	Separa perfiles poco frecuentes según cortes aleatorios en el espacio de variables.

Nota. Elaboración propia a partir de los resultados de priorización obtenidos en el notebook. Los modelos no son equivalentes entre sí; cada uno prioriza perfiles con una lógica diferente.

Figura 27 Perfiles de alta prioridad identificados por modelo

Nota. Elaboración propia a partir de los resultados de priorización del notebook. La figura compara la cantidad de perfiles de alta prioridad identificados por cada técnica.

La comparación muestra que GMM identifica más perfiles prioritarios, seguido por DBSCAN, Isolation Forest, HDBSCAN y K-Means. Esta diferencia no necesariamente es contradictoria, sino es la evidencia de que cada algoritmo mira una dimensión distinta del

comportamiento de los perfiles que se consideran como atípicos: volumen y señales, ausencia de densidad, incertidumbre de pertenencia o aislamiento del resto de los perfiles con agrupamiento.

8.8 Síntesis del capítulo

La etapa de puntuación permitió transformar los resultados del modelado no supervisado en criterios de priorización para el análisis de perfiles. A partir de los clusters, las señales tributarias, los perfiles clasificados como ruido, la incertidumbre probabilística y la detección directa de anomalías, se construyó una lectura integrada de los perfiles tributarios agregados.

Los puntajes y categorías de prioridad no reemplazan los procesos institucionales de control tributario ni constituyen evidencia concluyente de evasión fiscal. Su utilidad se encuentra en ordenar perfiles agregados según su nivel de interés analítico, de manera que puedan orientar revisiones posteriores, cruces con información adicional o análisis más detallados por parte de la administración tributaria.

El aporte principal de este enfoque consiste en no depender de un único modelo ni de una sola métrica de clasificación, ni encontrar el mejor modelo de agrupación para evasión. Por el contrario, se integraron distintas perspectivas: segmentación por similitud estadística, señales ubicadas en el tramo superior de variables relevantes, baja integración a zonas densas, incertidumbre de pertenencia y atipicidad directa. Esta lectura conjunta permite interpretar los resultados con mayor solidez y sirve como base para el análisis sectorial y territorial desarrollado en el capítulo siguiente.

CAPÍTULO 9

INTERPRETACIÓN, DISCUSIÓN Y CONCLUSIONES DEL MODELO

El presente capítulo hace la interpretación de los resultados obtenidos en las etapas de modelado, como son la puntuación y priorización. La gran diferencia con los capítulos anteriores es que el énfasis no está en ejecutar otros algoritmos, sino más bien poder leer los hallazgos de manera más integral y valorar cuan útiles son para llegar al objetivo general del proyecto.

La interpretación se mantiene dentro de los límites planteados para el estudio: los resultados se refieren a perfiles agregados por año, provincia y sector económico; no identifican contribuyentes individuales ni determinan evasión fiscal. El valor real del presente trabajo se encuentra en ofrecer señales de interés analítico, orientar preguntas de revisión y proponer criterios de priorización para apoyar investigaciones futuras o procesos complementarios de control tributario.

9.1 Lectura integrada de los modelos aplicados

Los modelos aplicados no entregan una única respuesta, sino varias lecturas complementarias del mismo fenómeno. K-Means organiza los perfiles en segmentos de comportamiento similar; DBSCAN y HDBSCAN identifican perfiles con baja integración a zonas densas; GMM permite observar perfiles con menor certeza de pertenencia a un único grupo; e Isolation Forest prioriza perfiles que se aíslan con mayor facilidad dentro del espacio de variables. Esta diversidad fortalece el análisis, y la lectura correcta es entendiendo que los resultados no se interpreten como si todos los modelos midieran exactamente lo mismo.

Desde la perspectiva tributaria, una anomalía estadística no es suficiente por sí sola para justificar una revisión complementaria ni para afirmar incumplimiento. Su valor consiste en orientar preguntas analíticas, activar comparaciones históricas, sugerir cruces con fuentes

adicionales o enfocar esta aplicación a estructuras granulares de contribuyentes. Por lo tanto, los perfiles priorizados deben entenderse como segmentos agregados que requieren revisión complementaria antes de cualquier decisión operativa.

Tabla 50

Síntesis interpretativa de los modelos aplicados

Modelo	Criterio de priorización	Perfiles de alta prioridad	Lectura analítica
K-Means	Cluster K-Means + señales por percentil 90	270	Prioriza perfiles de alta recaudación, alta dispersión y acumulación de señales.
DBSCAN	Ruido o puntos sin densidad suficiente	386	Identifica observaciones que no se integran a los grupos principales.
HDBSCAN	Ruido en agrupamiento jerárquico por densidad	344	Similar a DBSCAN, pero más flexible ante densidades diferentes.
GMM	Incertidumbre de pertenencia al cluster	467	Identifica perfiles con mayor ambigüedad de pertenencia a un único componente.
Isolation Forest	Aislamiento directo de anomalías	351	Separa perfiles poco frecuentes según cortes aleatorios en el espacio de variables.

Nota. Elaboración propia a partir de los resultados de los capítulos 7 y 8. La tabla resume la utilidad analítica de cada modelo dentro del enfoque no supervisado.

El resultado más defendible para la tesis no es seleccionar un único modelo como verdad final, sino usar los modelos como capas de evidencia analítica. Esta lectura es más prudente, más honesta con la naturaleza no supervisada del trabajo y más coherente con el uso de datos públicos agregados.

9.2 Análisis sectorial de perfiles priorizados

El análisis sectorial se realizó a partir de los perfiles clasificados como prioridad alta mediante K-Means. Los resultados muestran una mayor concentración en los sectores C y G, correspondientes a industrias manufactureras y comercio al por mayor y al por menor, seguidos por agricultura, construcción, actividades financieras, transporte, administración pública, explotación de minas, actividades inmobiliarias y actividades profesionales. Esta distribución

no implica que dichos sectores presenten incumplimiento tributario, sino que concentran perfiles agregados que tienen mayor interés para los análisis posteriores.

Tabla 51

Sectores con mayor cantidad de perfiles de alta prioridad

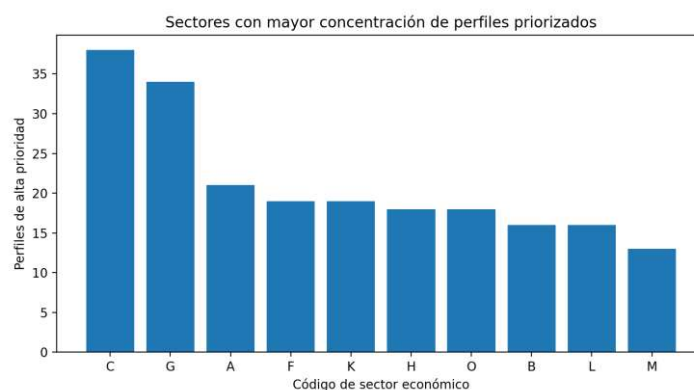
Código	Sector económico	Perfiles de alta prioridad
C	Industrias manufactureras	38
G	Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas	34
A	Agricultura, ganadería, silvicultura y pesca	21
F	Construcción	19
K	Actividades financieras y de seguros	19
H	Transporte y almacenamiento	18
O	Administración pública y defensa; planes de seguridad social obligatoria	18
B	Explotación de minas y canteras	16
L	Actividades inmobiliarias	16
M	Actividades profesionales, científicas y técnicas	13

Nota. Elaboración propia a partir de los perfiles clasificados como prioridad alta por K-Means. Los códigos

corresponden a secciones de actividad económica utilizadas en la información del SRI.

La presencia de sectores como manufactura, comercio, agricultura, construcción y actividades financieras es coherente con su peso dentro de la actividad económica y tributaria del país. Precisamente por su importancia recaudatoria, estos sectores ofrecen un espacio más amplio para observar variaciones, dispersión o comportamientos diferenciados entre perfiles agregados. Es decir, el hallazgo no señala malas prácticas tributarias sectoriales, sino que identifica ámbitos donde la priorización analítica puede resultar más útil por su carga tributaria en sí.

Figura 28 Sectores con mayor concentración de perfiles priorizados



Nota. Elaboración propia a partir de los resultados de priorización K-Means. La figura presenta los diez sectores con mayor número de perfiles de alta prioridad.

9.3 Análisis territorial de perfiles priorizados

En el caso análisis territorial se realizó a partir de los perfiles clasificados como prioridad alta mediante K-Means. Los resultados muestran una mayor concentración en Guayas y Pichincha, seguidas por Azuay, Manabí, El Oro, Tungurahua y Zamora Chinchipe. Esta distribución no debe interpretarse como evidencia de mayor evasión fiscal en dichas provincias, sino como una concentración de perfiles agregados de interés analítico dentro de los territorios con mayor peso económico y tributario.

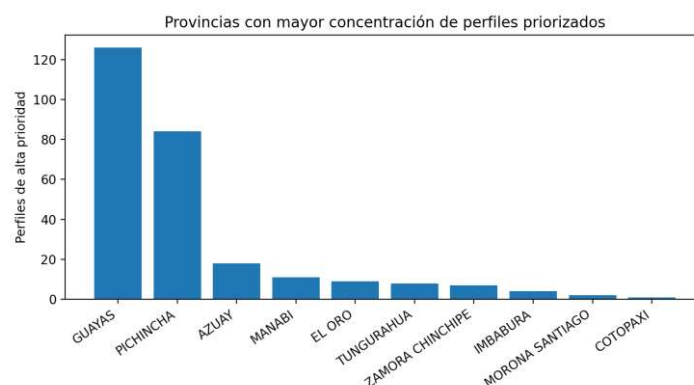
Tabla 52

Provincias con mayor cantidad de perfiles de alta prioridad

Provincia	Perfiles de alta prioridad
GUAYAS	126
PICHINCHA	84
AZUAY	18
MANABI	11
EL ORO	9
TUNGURAHUA	8
ZAMORA CHINCHIPE	7
IMBABURA	4
MORONA SANTIAGO	2
COTOPAXI	1

Nota. Elaboración propia a partir de los perfiles clasificados como prioridad alta por K-Means.

Figura 29 Provincias con mayor concentración de perfiles priorizados



Nota. Elaboración propia a partir de los resultados de priorización K-Means. La figura presenta las diez provincias con mayor número de perfiles de alta prioridad.

La concentración de perfiles priorizados en Guayas y Pichincha es de esperarse con el peso económico y recaudatorio de ambas provincias. En este sentido, una mayor cantidad de perfiles de alta prioridad puede responder al volumen de actividad económica registrada y no

necesariamente a una mayor presencia de incumplimiento tributario. Es por eso, el resultado debe leerse como una guía para orientar revisiones analíticas territoriales, especialmente en provincias donde la magnitud de la actividad económica hace más relevante la focalización de recursos de análisis.

9.4 Utilidad del puntaje de prioridad analítica como herramienta de apoyo

El puntaje de prioridad analítica construido en la investigación es útil como herramienta de apoyo para ordenar perfiles agregados y orientar las revisiones que optimizan recursos. Su principal ventaja es que permite resumir información compleja en categorías interpretables de priorización, facilitando la identificación de segmentos que combinan alta recaudación, dispersión, concentración temporal, volatilidad o comportamientos atípicos respecto del patrón predominante.

La utilidad práctica del puntaje se encuentra en su capacidad para enfocar la atención en sectores, provincias y combinaciones año-provincia-sector que presentan mayor interés analítico. En un contexto donde los recursos de análisis y control son limitados, una herramienta de priorización puede ayudar a definir por dónde iniciar una revisión más detallada. Pero se hace necesario aclarar que el puntaje no sustituye los criterios técnicos institucionales, la evidencia documental ni la decisión humana; funciona únicamente como un filtro inicial para delimitar segmentos que requieren revisión complementaria.

En una aplicación institucional, el puntaje de prioridad analítica serviría como una primera capa de monitoreo macro. El flujo recomendado sería: identificación de segmentos agregados prioritarios, revisión analítica, contraste con información complementaria y decisión humana sobre la necesidad de profundizar el análisis. Bajo este esquema, el modelo no automatiza auditorías ni reemplaza criterios jurídicos; únicamente ayuda a ordenar la atención hacia segmentos donde los datos muestran comportamientos menos esperados.

El uso de estos resultados debería complementarse con fuentes adicionales autorizadas, como información de catastro tributario, comprobantes electrónicos, registros societarios, comercio exterior, antecedentes de control u otras bases internas disponibles para la administración tributaria. Esta interacción puede construir una interpretación más amplia, reducir falsos positivos y diferenciar comportamientos coyunturales de patrones persistentes.

9.5 Limitaciones del análisis

Tabla 53

Limitaciones principales del proceso de modelado e interpretación

Limitación	Descripción	Tratamiento en la investigación
Datos agregados	La información se analiza por año, provincia y sector, no por contribuyente individual.	Interpretar resultados como señales agregadas y no como evidencia directa.
Ausencia de etiquetas	No existen etiquetas institucionales de evasión confirmada para entrenar o validar modelos supervisados.	Utilizar aprendizaje no supervisado y evitar afirmaciones concluyentes.
Sensibilidad a parámetros	Modelos como DBSCAN, HDBSCAN e Isolation Forest dependen de parámetros definidos por el investigador.	Reportar los criterios usados y comparar resultados entre modelos.
Diferencias de escala	Los perfiles de provincias y sectores grandes pueden dominar algunas señales.	Estandarizar variables y complementar con medidas relativas.
Interpretación tributaria limitada	Una anomalía estadística puede tener explicaciones económicas legítimas.	Requerir análisis contextual y fuentes complementarias antes de cualquier conclusión operativa.

Nota. Elaboración propia a partir de la revisión metodológica de los resultados. Las limitaciones no invalidan el análisis, pero delimitan el alcance de sus conclusiones.

La principal limitación metodológica del estudio es que los datos disponibles son públicos y agregados. Esta condición impide evaluar contribuyentes individuales, confirmar causalidad o validar el modelo frente a registros institucionales de evasión fiscal confirmada. La base analizada no incorpora RUC, expedientes tributarios individuales, comprobantes electrónicos detallados ni etiquetas institucionales de incumplimiento tributario.

A pesar de la limitación, no se invalida el estudio; por el contrario, define su alcance real. La investigación propone una metodología reproducible, prudente y compatible con restricciones legales de acceso a datos. En consecuencia, los resultados deben leerse como

señales de priorización macroanalítica y no como conclusiones definitivas sobre incumplimiento tributario.

9.6 Conclusiones del modelado

El proceso de modelado permitió identificar perfiles tributarios agregados con comportamientos diferenciados dentro de la recaudación de sociedades en Ecuador. La construcción de perfiles por año, provincia y sector económico ofreció una base de análisis adecuada para trabajar con datos públicos, mantener la comparabilidad territorial y sectorial, y evitar la identificación de contribuyentes individuales.

Los modelos aplicados aportaron lecturas complementarias. K-Means permitió segmentar perfiles según similitud estadística; DBSCAN y HDBSCAN identificaron perfiles con baja integración a zonas densas; GMM permitió observar perfiles con menor certeza de pertenencia a un único componente; e Isolation Forest priorizó observaciones con mayor grado de atipicidad. La integración de estas técnicas fortaleció la lectura analítica del fenómeno, sin convertir los resultados en evidencia concluyente de evasión fiscal.

La etapa de puntuación y priorización evidenció la existencia de un subconjunto de perfiles que acumula señales de interés analítico. Estos perfiles se concentran principalmente en sectores y provincias de alta relevancia económica, lo cual debe interpretarse como una oportunidad para orientar revisiones agregadas más detalladas y no como una afirmación de incumplimiento tributario.

En conclusión, el modelo desarrollado cumple una función de priorización macroanalítica. Su valor se encuentra en ordenar segmentos año-provincia-sector según patrones estadísticos atípicos, facilitando la focalización de análisis posteriores. No obstante, cualquier decisión operativa requeriría información adicional, validación institucional y revisión humana especializada.

9.7 Recomendaciones y trabajo futuro

Se recomienda mantener el enfoque de priorización como una herramienta exploratoria y no como un mecanismo automático de selección de contribuyentes o determinación tributaria. Los resultados del modelo deben utilizarse como apoyo para ordenar perfiles agregados, orientar preguntas analíticas y focalizar revisiones posteriores, siempre bajo supervisión técnica y decisión humana.

Como trabajo futuro, se recomienda ampliar la validación mediante análisis de sensibilidad de hiperparámetros, comparando distintos valores de k en K-Means, ϵ en DBSCAN, tamaños mínimos de cluster en HDBSCAN, número de componentes en GMM y niveles de contaminación en Isolation Forest. Esta evaluación permitiría determinar si los perfiles priorizados se mantienen estables frente a cambios razonables de parametrización.

También se recomienda evaluar la estabilidad temporal de los perfiles mediante ventanas anuales o subperiodos, con el fin de diferenciar comportamientos coyunturales de patrones persistentes. Esta validación ayudaría a identificar si los segmentos priorizados aparecen de forma aislada en un período específico o si mantienen señales similares a lo largo del tiempo.

En una implementación institucional, el modelo podría integrarse en un tablero de monitoreo periódico que actualice los perfiles con nueva información pública y, cuando exista autorización legal, con fuentes internas de mayor granularidad, como comprobantes electrónicos, catastro tributario, registros societarios, comercio exterior o antecedentes de control. Estas fuentes permitirían contrastar los perfiles agregados con información complementaria y reducir el riesgo de falsos positivos.

Finalmente, se recomienda que cualquier uso operativo del modelo conserve una lógica de apoyo y no de sustitución de los procesos institucionales. La decisión final debería permanecer en manos de analistas y funcionarios responsables, utilizando el puntaje de

prioridad analítica como una herramienta para priorizar segmentos, no como un mecanismo automático de determinación de veasión tributaria.

Referencias

- 2.7. *Novelty and Outlier Detection*. (s. f.). Scikit-Learn. Recuperado 15 de julio de 2026, de https://scikit-learn/stable/modules/outlier_detection.html
- 2020 *Manual-gestion-riesgos_CIAT-SII-FMI*. (s. f.).
- AI Principles Overview*. (s. f.-a). Recuperado 15 de julio de 2026, de <https://oecd.ai/en/principles>
- AI Principles Overview*. (s. f.-b). Recuperado 15 de julio de 2026, de <https://oecd.ai/en/principles>
- Balderas-Díaz, S., & Guerrero-Contreras, G. (s. f.). *V JORNADAS DE INVESTIGACIÓN PREDOCTORAL EN INGENIERÍA INFORMÁTICA FIFTH DOCTORAL CONSORTIUM IN COMPUTER SCIENCE*.
- Barrezueta, H. D. P. (s. f.). *DIRECTOR DEL REGISTRO OFICIAL*.
- Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. En J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160-172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14
- Censos, I. N. de E. y. (s. f.). *Registro Estadístico de Empresas – REEM 2024 (Definitivo) y 2025 (Provisional)*. Instituto Nacional de Estadística y Censos. Recuperado 15 de julio de 2026, de <https://www.ecuadorencifras.gob.ec/directoriodeempresas/>
- Chen Jinyin, & Yang Dongyong. (2018). (PDF) A k -Deviation Density Based Clustering Algorithm. *ResearchGate*. <https://doi.org/10.1155/2018/3742048>
- Codigo_Tributario_20_junio_2023*. (s. f.).
- Consensos y conflictos en la política tributaria de América Latina | CEPAL*. (s. f.). Recuperado 7 de junio de 2026, de <https://www.cepal.org/es/publicaciones/41048-consensos-conflictos-la-politica-tributaria-america-latina>

- Dabbura, I. (2022, septiembre 27). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. *TDS Archive*. <https://medium.com/data-science/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>
- Delua, J. (2024, marzo 12). *Supervised vs. Unsupervised Learning: What's the Difference?* | IBM. <https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning>
- Dhanesha, P., & Mehta, D. (2025). Uncovering Hidden Frauds: Isolation Forest-Based Anomaly Detection in Credit Card Transactions. En M. S. Kaiser, J. Xie, & V. S. Rathore (Eds.), *Intelligent Strategies for ICT* (pp. 39-51). Springer Nature. https://doi.org/10.1007/978-981-96-4145-1_4
- Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. (s. f.). Recuperado 15 de julio de 2026, de <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Forke, C.-M., & Tropmann-Frick, M. (2021). Feature Engineering Techniques and Spatio-Temporal Data Processing. *Datenbank-Spektrum*, 21(3), 237-244. <https://doi.org/10.1007/s13222-021-00391-x>
- González, M. S. (s. f.). *APLICACIÓN DE MODELO DE DETECCIÓN DE ANOMALIAS ISOLATION FOREST PARA LA DETECCIÓN DE RECLAMOS FRAUDULENTOS EN UNA COMPAÑÍA DEL SECTOR DE MEDICINA PREPAGADA*.
- Hernández, C., Grijalva, F., Parra, C., Riofrío, D., & Vega-Sánchez, J. (2025). Unsupervised Learning for Fraud Detection in Cooperative Transactions: 9th Ecuador Technical Chapters Meeting, ETCM 2025. *ETCM 2025 - 9th Ecuador Technical Chapters Meeting, ETCM 2025 - 9th Ecuador Technical Chapters Meeting*. <https://doi.org/10.1109/ETCM67548.2025.11304315>
- International Encyclopedia of Statistical Science. (2025). Principal Component Analysis. En *ResearchGate*. https://doi.org/10.1007/978-3-662-69359-9_483

IsolationForest. (s. f.). Scikit-Learn. Recuperado 15 de julio de 2026, de <https://scikit-learn/stable/modules/generated/sklearn.ensemble.IsolationForest.html>

Judith, C., Uzoma, J., Ugochi, Q., & Churchill, C. (2023). *Machine Learning Algorithms for Detecting Tax Fraud: Application and Challenges in Nigeria*.

Kavlakoglu, J. M. P. D., Eda. (2024, enero 5). *¿Qué es la reducción de la dimensionalidad?* | IBM. <https://www.ibm.com/es-es/think/topics/dimensionality-reduction>

Koukaras, P., & Tjortjis, C. (2025). Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. *AI*, 6(10), 257. <https://doi.org/10.3390/ai6100257>

Matriano, D. M. T. (2024). *Green Finance and How Sustainability is Shaping the Investment Strategies*. 12(12).

Mohd Padil, H., Kasim, E. S., Muda, S., Ismail, N., & Md Zin, N. (2021). Financial literacy and awareness of investment scams among university students. *Journal of Financial Crime*, 29(1), 355-367. <https://doi.org/10.1108/JFC-01-2021-0012>

Nelli, F. (2023). *Python Data Analytics: With Pandas, NumPy, and Matplotlib*. Apress. <https://doi.org/10.1007/978-1-4842-9532-8>

Noble, J. (2025a, septiembre 27). *What is a Gaussian Mixture Model?* | IBM. <https://www.ibm.com/think/topics/gaussian-mixture-model>

Noble, J. (2025b, septiembre 27). *What is a Gaussian Mixture Model?* | IBM. <https://www.ibm.com/think/topics/gaussian-mixture-model>

(PDF) LQG-Based Virtual Inertial Control of Islanded Microgrid Load Frequency Control and DoS Attack Vulnerability Analysis. (2026). *ResearchGate*. <https://doi.org/10.1109/ACCESS.2023.3271012>

¿Qué es el aprendizaje no supervisado? Google Cloud | Google Cloud. (s. f.). Recuperado 15 de julio de 2026, de <https://cloud.google.com/discover/what-is-unsupervised-learning>

Radosław J. WOŹNIAK. (2025). *Unsupervised machine learning in financial anomaly detection: Clustering algorithms vs. Dedicated methods*.

<https://przegladteleinformatyczny.publisherspanel.com/article/548748/en>

RAE. (s. f.). *Definición de evasión fiscal—Diccionario panhispánico del español jurídico—RAE*. Diccionario panhispánico del español jurídico - Real Academia Española.

Recuperado 7 de junio de 2026, de <https://dpej.rae.es/lema/evasi%C3%B3n-fiscal>

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9(1), 59. <https://doi.org/10.1186/s40537-022-00592-5>

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3), 19:1-19:21. <https://doi.org/10.1145/3068335>

StandardScaler. (s. f.). Scikit-Learn. Recuperado 15 de julio de 2026, de <https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>

The hdbscan Clustering Library—Hdbscan 0.8.1 documentation. (s. f.). Recuperado 15 de julio de 2026, de <https://hdbscan.readthedocs.io/en/latest/>

Vallejo-Vega, M. J., Pinda-Guanolema, B. R., Romero-Fernández, A. J., Vallejo-Vega, M. J., Pinda-Guanolema, B. R., & Romero-Fernández, A. J. (2025). La evasión fiscal en el Ecuador. *Noesis*, 7(ESP1), 830-842. <https://doi.org/10.35381/noesisin.v7i1.448>

wgaleas. (2021, mayo 12). La pandemia por el COVID-19 generó una caída en el PIB de 6,4% de marzo a diciembre de 2020. *Banco Central del Ecuador*. <https://www.bce.fin.ec/la-pandemia-por-el-covid-19-genero-una-caida-en-el-pib-de-6-4-de-marzo-a-diciembre-de-2020/>

What is Exploratory Data Analysis? | IBM. (s. f.). Recuperado 15 de julio de 2026, de <https://www.ibm.com/think/topics/exploratory-data-analysis>

Winland, E. K., Vanna. (2024, junio 26). *What is k-means clustering?* | IBM.

<https://www.ibm.com/think/topics/k-means-clustering>

Xie, J., & Wan, J. (2023). Digital Twin Four-Dimension Fusion Modeling Method Design and Application to the Discrete Manufacturing Line. *Big Data and Cognitive Computing*, 7(2), 89. <https://doi.org/10.3390/bdcc7020089>