PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

TRABAJO DE TITULACIÓN

Elaboración de un modelo de aprendizaje automático no supervisado para describir patrones de ingresos tributarios en el Ecuador, utilizando datos abiertos gubernamentales durante el período 2017–2025

Autor: Francisco Javier Rodríguez Narváez

Facultad: Hábitat, Infraestructura y Creatividad

Programa: Sistemas de Información con mención en Data Science

Director: Charles Édisson Escobar Terán

Fecha: Agosto 2026

Agradecimiento

Agradezco a Dios, a mi familia y a mis profesores por acompañarme, apoyarme y orientarme a lo largo de este proceso. Su presencia y guía fueron fundamentales para culminar este reto académico.

Contenido

Resumen ejecutivo 8

Capítulo 1. Introducción 10

1.1 Justificación 10

1.2 Alcance y delimitaciones 11

1.3 Planteamiento del problema 11

1.4 Preguntas de investigación 13

1.5 Objetivos 13

1.5.1 Objetivo general 13

1.5.2 Objetivos específicos 13

Capítulo 2. Marco teórico 15

2.1 Recaudación tributaria y financiamiento público en el Ecuador 15

2.2 Evolución histórica de la recaudación tributaria durante el período 2017–2025 16

2.3 Ventas declaradas y su relación con el análisis tributario 16

2.4 Datos abiertos gubernamentales 17

2.5 Fuentes de datos de Recaudación y Ventas-Compras del Servicio de Rentas Internas 17

2.6 Variables temporales, territoriales, económicas y tributarias 18

2.7 Calidad, limpieza, homologación e integración de datos 18

2.8 Aprendizaje automático no supervisado 19

2.9 Agrupamiento de datos o clustering 19

2.9.1 K-Means 19

2.9.2 K-medoids 20

2.9.3 DBSCAN 20

2.9.4 Agrupamiento jerárquico 20
2.10 Evaluación de modelos de agrupamiento 21
2.10.1 Coeficiente Silhouette 21
2.10.2 Índice Davies-Bouldin 21
2.10.3 Suma de errores cuadrados y método del codo 21
2.10.4 Estabilidad e interpretabilidad de los clústeres 22
2.11 Reducción de dimensionalidad mediante componentes principales 22
2.12 Visualización de datos y dashboards interactivos 22
2.13 Analítica tributaria para apoyar la toma de decisiones fiscales 23
Capítulo 3. Metodología 24
3.1 Enfoque y diseño de investigación 24
3.2 Referencia metodológica: CRISP-DM 24
3.3 Relación entre CRISP-DM y el proyecto 25
3.4 Fuentes de información y unidad de análisis 25
3.5 Preparación, modelado y evaluación 25
3.6 Implementación del dashboard y reproducibilidad 26
3.7 Herramientas utilizadas 26
Capítulo 4. Desarrollo del proceso analítico integral aplicado a los datos abiertos tributarios del Ecuador correspondientes al período 2017–2025 28
4.1 Extracción de datos de Recaudación y Ventas-Compras 28
4.2 Preparación de los datos tributarios obtenidos 29
4.3 Modelado de patrones tributarios mediante técnicas de aprendizaje automático no supervisado 31

4.4 Evaluación de los modelos de agrupamiento 32

4.4.1 Comparación y selección del modelo 32

4.4.2 Resultados del modelo seleccionado 34

4.4.3 Validación externa institucional de la solución seleccionada 38

4.5 Implementación de un dashboard interactivo 39

4.6 Análisis de resultados 46

Capítulo 5. Conclusiones, limitaciones y recomendaciones 54

5.1 Conclusiones 54

5.2 Limitaciones 55

5.3 Recomendaciones 56

Referencias 57

Anexos 61

Anexo A. Relación entre notebooks y entregables 61

Anexo B. Archivos principales generados 61

Anexo C. Nota de reproducibilidad y alcance de los resultados 62

Índice de tablas

Tabla 1 *Relación entre CRISP-DM y el desarrollo del proyecto* 25
Tabla 2 *Resumen de extracción e integración inicial* 28
Tabla 3 *Bases analíticas preparadas* 29
Tabla 4 *Resumen comparativo de modelos candidatos* 32
Tabla 5 *Perfil resumido de clústeres* 35
Tabla 6 *Participación de registros y concentración monetaria por clúster* 36
Tabla 7 *Estructura funcional final del dashboard* 39
Tabla 8 *Indicadores principales del dashboard para 2025* 42
Tabla 9 *Principales provincias por recaudación acumulada* 44
Tabla 10 *Principales sectores por recaudación y recaudación por registro* 45
Tabla 11 *Indicadores complementarios para la interpretación del clúster 0* 48
Tabla 12 *Trazabilidad de notebooks y entregables principales* 61

Índice de figuras

Figura 1 *Varianza acumulada explicada por PCA en las versiones completa y reducida* 30
Figura 2 *Comparación del coeficiente Silhouette de K-Means según número de clústeres* 33
Figura 3 *Distribución de la solución K-Means reducida con tres clústeres en el plano PCA* 35
Figura 4 *Distribución de registros y participación monetaria por clúster* 37
Figura 5 *Evolución anual de la participación de registros por clúster* 37
Figura 6 *Valor recaudado por clúster en las principales provincias* 40
Figura 7 *Valor recaudado por clúster en los principales sectores económicos* 41
Figura 8 *Evolución anual de la recaudación y variación interanual* 43
Figura 9 *Participación de registros frente a participación en recaudación por clúster* 44
Figura 10 *Concentración territorial de la recaudación en las diez principales provincias* 45
Figura 11 *Composición porcentual de los clústeres en los principales sectores económicos* 52

Resumen ejecutivo

Este trabajo desarrolla un proceso analítico integral para describir patrones de ingresos tributarios en el Ecuador mediante técnicas de aprendizaje automático no supervisado. La investigación utiliza datos abiertos del Servicio de Rentas Internas correspondientes al período 2017–2025, específicamente las fuentes de Recaudación y Ventas-Compras. El punto de partida fue la dispersión de la información en archivos anuales y en fuentes con estructuras y niveles de detalle que debían comprenderse antes de cualquier integración.

La ruta metodológica se organizó con referencia a CRISP-DM, sin aplicarla como una plantilla rígida. El proceso quedó documentado en cinco notebooks: extracción, diagnóstico e integración preliminar; preparación de variables; modelado no supervisado y comparación de alternativas; consolidación de resultados; e implementación del dashboard en Streamlit. Esta separación mantiene la trazabilidad entre fuentes, transformaciones, modelos, tablas finales y visualizaciones.

El modelo seleccionado fue K-Means aplicado a una base reducida y escalada con tres clústeres. La decisión combinó métricas internas, posibilidad de aplicación sobre las 391.774 observaciones, costo computacional, estabilidad práctica e interpretabilidad. Los perfiles obtenidos fueron: ventas bajas y compras relativamente altas; actividad frecuente de menor magnitud; y alta magnitud económica y comercio exterior. El segundo perfil reúne el 80,78 % de los registros, mientras que el tercero concentra el 96,51 % de la recaudación.

El producto final incorpora un dashboard multipágina en Streamlit que consume directamente las tablas preparadas por los notebooks. Para 2025 se observaron 43.302 registros agregados y una recaudación de $ 20.900.488.321,68, equivalente a un crecimiento interanual de 4,81 %. En el período completo, Pichincha y Guayas acumulan el 82,71 % de la recaudación, y

el sector de comercio presenta el mayor monto recaudado. Estos resultados muestran que frecuencia, magnitud económica, territorio y actividad sectorial aportan lecturas complementarias.

El alcance del estudio es académico, descriptivo y exploratorio. Las observaciones corresponden a combinaciones agregadas de tiempo, territorio y sector económico; por ello, los resultados no identifican contribuyentes, no estiman evasión, no clasifican riesgo individual y no permiten establecer relaciones causales. La contribución del trabajo consiste en ofrecer una ruta reproducible para transformar datos abiertos tributarios en bases analíticas, modelos comparados, perfiles interpretables y visualizaciones que apoyan el análisis fiscal agregado.

Capítulo 1. Introducción

1.1 Justificación

El proyecto se justifica porque trabaja con datos públicos que poseen valor potencial para el análisis fiscal, pero que necesitan un proceso previo de organización. En un contexto de apertura gubernamental, los datasets no solo deben estar disponibles; también deben ser interpretables, trazables y adecuados al uso previsto. La literatura sobre datos abiertos ha mostrado que la apertura puede favorecer la transparencia y la reutilización de la información, aunque su aprovechamiento puede verse limitado por barreras relacionadas con los formatos, los metadatos, la fragmentación y las capacidades técnicas (Janssen et al., 2012).

Desde una perspectiva técnica, el trabajo aporta una comparación documentada de modelos no supervisados. Esta decisión evita asumir que un único algoritmo resulta adecuado para cualquier estructura de datos. K-Means permite trabajar de manera eficiente con conjuntos de datos amplios; K-medoids utiliza observaciones reales como representantes de los grupos y presenta mayor robustez frente a valores extremos; DBSCAN permite identificar grupos por densidad y reconocer observaciones como ruido; y el método de Ward permite construir una estructura jerárquica de agrupamiento (MacQueen, 1967; Kaufman & Rousseeuw, 1990; Ester et al., 1996; Ward, 1963). Comparar estos enfoques permite seleccionar el modelo a partir de varios criterios y no únicamente de una métrica.

Desde el punto de vista aplicado, la investigación contribuye con una ruta reproducible. Cada decisión relevante queda conectada con archivos, variables, transformaciones, parámetros y resultados. Este énfasis es coherente con recomendaciones recientes para la analítica tributaria, donde se destaca la importancia de trabajar con datos adecuados, procesos documentados y

criterios claros de revisión y supervisión para apoyar la toma de decisiones (Aslett, González, Hamilton, & Pecho, 2024; OECD, 2024).

El dashboard también justifica el proyecto porque traduce resultados técnicos en una forma de consulta más accesible. La visualización no sustituye el análisis, pero ayuda a reconocer comparaciones temporales, territoriales y sectoriales. Munzner (2014) plantea que una visualización debe responder qué datos se muestran, para qué tarea se examinan y cómo conviene representarlos. Ese principio orienta la implementación en Streamlit.

1.2 Alcance y delimitaciones

El alcance del estudio es descriptivo, exploratorio y académico. Se trabaja con datos abiertos agregados del SRI, no con registros individuales de contribuyentes. Por tanto, los resultados no deben utilizarse para inferir conducta individual, riesgo de evasión, incumplimiento tributario ni causalidad económica.

La delimitación temporal corresponde al período 2017–2025. La delimitación temática se concentra en las fuentes de Recaudación y Ventas-Compras. La delimitación metodológica se orienta al aprendizaje automático no supervisado y, dentro de este campo, a algoritmos de agrupamiento evaluados mediante métricas internas y criterios interpretativos.

El dashboard forma parte del producto como prototipo de comunicación visual. No se plantea como plataforma institucional en producción. Su función es apoyar la exploración de resultados y documentar una fase de despliegue analítico reproducible.

1.3 Planteamiento del problema

El Servicio de Rentas Internas publica datasets de Recaudación y de Ventas-Compras que permiten estudiar información tributaria agregada del Ecuador para el período 2017–2025 (Servicio de Rentas Internas, s. f.). Sin embargo, la existencia pública de esos archivos no

resuelve por sí misma el problema analítico. Los datos están organizados por año y por fuente, con estructuras, campos, categorías y niveles de detalle que deben revisarse antes de construir una base común.

El problema no se limita a descargar archivos. Para que la información pueda sostener un proceso de modelado, primero debe existir un procedimiento claro de inventario, diagnóstico, limpieza, homologación e integración. De lo contrario, los resultados podrían depender de decisiones aisladas o de uniones forzadas entre variables que no comparten la misma granularidad. Esta preocupación coincide con la literatura sobre datos abiertos y calidad de datos, que advierte que la disponibilidad no equivale automáticamente a reutilización válida ni a conocimiento confiable (Janssen et al., 2012; Wang & Strong, 1996).

A la dificultad de preparación se suma una decisión metodológica relevante. Los clústeres no están definidos previamente en las fuentes, por lo que no existe una etiqueta externa que permita saber de antemano qué algoritmo agrupa mejor los datos. K-Means, K-medoids, DBSCAN y el agrupamiento jerárquico parten de supuestos distintos sobre distancia, densidad, centralidad y estructura. Por esa razón, la selección del modelo requiere combinar métricas internas con estabilidad, interpretabilidad y utilidad dentro del contexto tributario.

Finalmente, los resultados deben comunicarse de forma comprensible. Un proceso técnico puede perder valor si queda reducido a tablas dispersas, notebooks o archivos intermedios. Por ello, el trabajo incluye la preparación de un dashboard en Streamlit que reúna indicadores, patrones temporales, lectura territorial, análisis sectorial y perfiles de clústeres. Su propósito es facilitar la exploración agregada de la evidencia, sin automatizar decisiones ni convertir los clústeres en diagnósticos individuales.

1.4 Preguntas de investigación

La pregunta general que orienta el estudio es: ¿cómo puede desarrollarse un proceso analítico integral que permita extraer, preparar, modelar, evaluar y visualizar datos abiertos tributarios del Ecuador para identificar patrones agregados útiles para el análisis fiscal?

De esta pregunta se derivan las siguientes preguntas específicas:

1. ¿Cómo organizar y consolidar los archivos de Recaudación y Ventas-Compras publicados por el SRI para el período 2017–2025?
2. ¿Qué decisiones de limpieza, homologación, transformación y normalización son necesarias para construir una base analítica reproducible?
3. ¿Qué técnicas de aprendizaje automático no supervisado permiten describir de mejor manera la estructura de los datos disponibles?
4. ¿Qué criterios técnicos e interpretativos deben considerarse para seleccionar el modelo de agrupamiento final?
5. ¿Cómo comunicar los resultados mediante un dashboard interactivo que mantenga el enfoque descriptivo y agregado del estudio?

1.5 Objetivos

1.5.1 Objetivo general

Desarrollar un proceso analítico integral para extraer, preparar, modelar, evaluar y visualizar los datos abiertos tributarios del Ecuador correspondientes al período 2017–2025.

1.5.2 Objetivos específicos

1. Extraer datos de Recaudación y Ventas-Compras desde el repositorio de datos abiertos del Ecuador publicado por el Servicio de Rentas Internas, correspondientes al período 2017–2025.

2. Preparar los datos tributarios obtenidos, considerando la revisión de columnas disponibles, calidad, valores faltantes, duplicados, categorías, transformaciones y normalización necesarias para el modelado.
3. Modelar patrones tributarios mediante técnicas de aprendizaje automático no supervisado, de acuerdo con las características, la escala y la granularidad de los datos disponibles.
4. Evaluar los modelos de agrupamiento a partir de métricas internas, estabilidad, interpretabilidad y utilidad dentro del contexto tributario.
5. Implementar un dashboard interactivo que presente indicadores, patrones temporales y territoriales, perfiles de agrupamiento y métricas de evaluación para apoyar el análisis y la toma de decisiones fiscales.

Capítulo 2. Marco teórico

El marco teórico se organiza a partir de la estructura definida para el proyecto y conecta el contexto tributario con la preparación de datos, los métodos de agrupamiento y la visualización. Su propósito no es reunir definiciones aisladas, sino explicar qué conocimiento previo respalda las decisiones metodológicas y cómo debe interpretarse el alcance de los resultados.

La revisión de literatura tuvo un enfoque semisistemático. Se combinaron búsquedas temáticas, rastreo de referencias y consulta de fuentes institucionales. Se priorizaron publicaciones recientes para analítica y administración tributaria, y se conservaron obras seminales cuando introducen algoritmos o métricas que siguen vigentes. La selección consideró pertinencia, calidad académica, trazabilidad y aplicabilidad al problema; PRISMA 2020 se utilizó como orientación para ordenar el proceso, sin presentar la revisión como exhaustiva ni como metaanálisis (Page et al., 2021).

2.1 Recaudación tributaria y financiamiento público en el Ecuador

La recaudación tributaria constituye una fuente central de financiamiento público y refleja la interacción entre actividad económica, normativa, cumplimiento y capacidad administrativa. En este trabajo se la examina desde una perspectiva descriptiva, con base en montos agregados publicados por el SRI. La literatura internacional señala que la analítica puede ayudar a comprender la composición de los ingresos, orientar preguntas de gestión y mejorar la utilización de la información, siempre que exista gobierno de datos, trazabilidad y revisión humana (OECD, 2016, 2024; Aslett, González, Hamilton, & Pecho, 2024).

El análisis no equipara recaudación con bienestar, eficiencia fiscal ni cumplimiento individual. Un incremento del monto recaudado puede estar relacionado con múltiples factores

que no se identifican causalmente en estos datos. Por ello, los resultados se interpretan como patrones agregados y no como una evaluación normativa del sistema tributario.

2.2 Evolución histórica de la recaudación tributaria durante el período 2017–2025

El período 2017–2025 permite observar continuidad y cambios en nueve años consecutivos. La comparación temporal se realiza mediante niveles, participaciones y variaciones interanuales, sin atribuir automáticamente los cambios a una política o acontecimiento específico. Esta cautela es necesaria porque las series agregadas resumen efectos económicos, administrativos y normativos que requerirían diseños adicionales para ser explicados causalmente.

La dimensión temporal también permite comprobar si los perfiles de agrupamiento aparecen de manera persistente o si responden a un solo año. En el proyecto, la estabilidad de la presencia de los tres clústeres a lo largo del período aporta evidencia descriptiva para su interpretación, aunque no implica que las observaciones mantengan una trayectoria individual en el tiempo.

2.3 Ventas declaradas y su relación con el análisis tributario

Las ventas declaradas aportan una aproximación agregada a la magnitud de la actividad económica registrada en las fuentes. Su comparación con recaudación, compras, exportaciones e importaciones ayuda a distinguir perfiles de escala y composición. Sin embargo, las ventas no equivalen a utilidad, base imponible ni capacidad contributiva, y una razón entre recaudación y ventas no debe interpretarse de forma automática como cumplimiento o incumplimiento.

El valor analítico de Ventas-Compras radica en complementar la lectura tributaria con variables económicas observables en el mismo nivel de agregación. La integración solo es válida cuando año, mes, provincia, cantón y sector son compatibles entre fuentes; las dimensiones

específicas, como impuesto o tarifa, se conservan en vistas separadas cuando su combinación podría duplicar montos.

2.4 Datos abiertos gubernamentales

Los datos abiertos gubernamentales amplían las posibilidades de investigación y control social, pero la disponibilidad pública no garantiza reutilización válida. Janssen et al. (2012) advierten que la fragmentación, los formatos heterogéneos, la falta de metadatos y las capacidades técnicas pueden limitar los beneficios esperados. En consecuencia, la apertura debe acompañarse de documentación, control de versiones y comprensión del significado de cada campo.

La reproducibilidad exige conservar los archivos originales, registrar la fecha de descarga, documentar transformaciones y separar claramente datos fuente, datos intermedios, resultados del modelo y productos de visualización. Esta trazabilidad permite revisar decisiones y evita que el dashboard se convierta en una fuente independiente de resultados.

2.5 Fuentes de datos de Recaudación y Ventas-Compras del Servicio de Rentas Internas

El SRI publica los datasets de Recaudación y Ventas-Compras utilizados en la investigación (Servicio de Rentas Internas, s. f.). Los archivos de Recaudación contienen dimensiones temporales, territoriales, sectoriales y tributarias; los de Ventas-Compras incorporan montos económicos agregados provenientes de declaraciones. La coincidencia parcial entre dimensiones permite construir una base común, pero no elimina la obligación de revisar códigos, nombres, tipos y granularidad.

El estudio trabaja con nueve archivos anuales de cada fuente, correspondientes a 2017–2025. La unidad de análisis final no representa personas ni empresas: corresponde a registros

agregados definidos por combinaciones compatibles de tiempo, territorio y actividad económica. Esta precisión condiciona todas las interpretaciones posteriores.

2.6 Variables temporales, territoriales, económicas y tributarias

Las variables temporales —año y mes— permiten estudiar evolución y variación; las territoriales —provincia y cantón— permiten examinar concentración espacial; el sector económico aporta una dimensión de actividad; y los montos de recaudación, ventas, compras, exportaciones e importaciones describen magnitud. Las variables derivadas, como razones y transformaciones logarítmicas, se incorporan para comparar escalas y relaciones, no para sustituir el significado de las variables originales.

La selección de variables debe evitar redundancia excesiva y dominancia por escala. Una matriz de clustering basada únicamente en montos brutos tendería a agrupar por tamaño, mientras que razones extremas podrían dominar la distancia. Por ello, la preparación combina tratamiento de valores extremos, selección conceptual y escalamiento.

2.7 Calidad, limpieza, homologación e integración de datos

La calidad se entiende como adecuación al uso analítico. Wang y Strong (1996) distinguen dimensiones intrínsecas, contextuales, representacionales y de accesibilidad; por tanto, un dato puede ser numéricamente correcto y aun así resultar inadecuado si su unidad, categoría o nivel de detalle no son comparables. Batini et al. (2009) plantean que la mejora de calidad requiere diagnóstico, medición y acciones documentadas.

En esta investigación se revisan valores faltantes, duplicados, tipos de datos, categorías, ceros, valores extremos, cambios de codificación y granularidad. La homologación no significa forzar equivalencias: cuando dos variables no comparten significado o nivel de detalle, se

mantienen separadas. Esta decisión reduce el riesgo de duplicar montos o construir relaciones artificiales.

2.8 Aprendizaje automático no supervisado

El aprendizaje automático no supervisado busca estructuras en datos que no poseen una variable objetivo conocida. En clustering, las observaciones se agrupan según una noción de similitud definida por las variables, sus transformaciones y la distancia utilizada. No existe un algoritmo universalmente superior; la elección depende de escala, forma, ruido, dimensionalidad, tamaño de la base y finalidad analítica (Han et al., 2012; Xu & Wunsch, 2009).

Los clústeres son resultados del método aplicado y no categorías naturales preexistentes. Su utilidad depende de que puedan describirse, contrastarse y relacionarse con el propósito del estudio. Por ello, el análisis combina métricas internas con tamaño de grupos, estabilidad, costo computacional e interpretación tributaria.

2.9 Agrupamiento de datos o clustering

2.9.1 K-Means

En K-Means, la asignación se realiza según la proximidad a los centroides y el procedimiento busca reducir la dispersión intragrupo (MacQueen, 1967). Su principal ventaja es la escalabilidad, especialmente relevante para una base de cientos de miles de registros. Sin embargo, sus resultados pueden verse afectados por la selección inicial de los centroides, la presencia de valores extremos y la existencia de agrupaciones con formas, tamaños o densidades diferentes. Por estas razones, antes de interpretar los resultados es necesario escalar las variables, ejecutar el algoritmo con distintas inicializaciones y contrastar la solución obtenida mediante métricas e interpretación sustantiva (Han et al., 2012; Xu & Wunsch, 2009).

2.9.2 K-medoids

K-medoids utiliza observaciones reales como representantes de los grupos y puede ser más robusto frente a valores extremos (Kaufman & Rousseeuw, 1990). Las mejoras computacionales recientes reducen su costo, aunque continúa siendo más exigente que K-Means en bases voluminosas (Schubert & Rousseeuw, 2021). Por esta razón, en el proyecto se evalúa sobre una muestra reproducible.

2.9.3 DBSCAN

DBSCAN define los clústeres como regiones de alta densidad separadas por zonas menos densas y permite clasificar determinadas observaciones como ruido. La formación de los grupos depende principalmente de dos parámetros: la distancia de vecindad y el número mínimo de observaciones requerido para formar una región densa (Ester et al., 1996). Aunque este enfoque permite identificar estructuras no necesariamente esféricas, su aplicación puede resultar más compleja cuando existen densidades heterogéneas o un número elevado de variables, debido a que las medidas de distancia tienden a perder capacidad discriminante en espacios de alta dimensionalidad (Han et al., 2012; Xu & Wunsch, 2009).

2.9.4 Agrupamiento jerárquico

El agrupamiento jerárquico construye una secuencia de fusiones o divisiones entre observaciones o grupos, cuya estructura puede representarse mediante un dendrograma. En particular, el método de Ward busca que cada nueva unión produzca el menor incremento posible de la variabilidad interna de los grupos (Ward, 1963). Esta representación facilita la exploración de relaciones entre agrupaciones y posibles niveles de partición. No obstante, el cálculo y almacenamiento de las distancias entre observaciones puede requerir una cantidad considerable de memoria y tiempo de procesamiento conforme aumenta el tamaño de la base,

por lo que en conjuntos voluminosos suele ser necesario trabajar con muestras, agregaciones o estrategias computacionales complementarias (Han et al., 2012; Xu & Wunsch, 2009).

2.10 Evaluación de modelos de agrupamiento

La evaluación no supervisada requiere varios criterios porque no existe una etiqueta externa que determine la respuesta correcta. Las métricas internas describen propiedades geométricas de la solución, pero no garantizan por sí solas utilidad sustantiva. Una solución con una cifra favorable puede generar demasiados grupos, grupos muy pequeños o perfiles difíciles de explicar.

2.10.1 Coeficiente Silhouette

El coeficiente Silhouette compara la cohesión de una observación con su propio grupo y su separación respecto del grupo vecino (Rousseeuw, 1987). Valores más altos sugieren mayor separación promedio, aunque la interpretación depende de la forma de los clústeres, la distancia y la muestra utilizada.

2.10.2 Índice Davies-Bouldin

El índice Davies-Bouldin relaciona dispersión interna y separación entre grupos; valores menores indican, en principio, clústeres más compactos y diferenciados (Davies & Bouldin, 1979). Al igual que otras métricas, favorece ciertas geometrías y debe leerse junto con evidencia adicional.

2.10.3 Suma de errores cuadrados y método del codo

La suma de errores cuadrados mide la variación interna en K-Means y disminuye a medida que aumenta el número de clústeres. El método del codo busca un punto en el que la reducción adicional deja de ser sustantiva. No constituye una regla automática, porque el punto de inflexión puede ser ambiguo y debe contrastarse con métricas e interpretación.

2.10.4 Estabilidad e interpretabilidad de los clústeres

La estabilidad examina si la estructura se mantiene ante muestras o inicializaciones reproducibles. La interpretabilidad verifica si los grupos pueden describirse mediante variables relevantes y si esa descripción evita inferencias que los datos no sostienen. En el proyecto también se consideran tamaño, equilibrio, aplicabilidad a toda la base y utilidad para la comunicación visual.

2.11 Reducción de dimensionalidad mediante componentes principales

El análisis de componentes principales resume variabilidad en componentes no correlacionados y puede apoyar la exploración de matrices con múltiples variables (Jolliffe & Cadima, 2016). En esta investigación se utiliza como recurso diagnóstico y de visualización. Los componentes no sustituyen la interpretación de recaudación, ventas, compras o comercio exterior, ni se emplean para asignar significado automático a los grupos.

2.12 Visualización de datos y dashboards interactivos

La visualización debe responder a los datos disponibles, a las tareas del usuario y a la forma de representación más adecuada (Munzner, 2014). En el dashboard, las tarjetas resumen niveles y cambios; las series temporales muestran evolución; los mapas y rankings permiten comparar territorios; y los perfiles facilitan la lectura de los clústeres.

El dashboard se concibe como una capa de consulta, no como un motor de modelado. Consume tablas finales generadas por los notebooks y aplica filtros y agregaciones de presentación. Esta separación protege la reproducibilidad y evita que una interacción visual modifique las etiquetas o los resultados del modelo.

2.13 Analítica tributaria para apoyar la toma de decisiones fiscales

La analítica tributaria puede apoyar la comprensión de riesgos, la priorización de preguntas y la organización de evidencia, pero su uso responsable exige datos pertinentes, gobernanza y supervisión humana (OECD, 2016, 2024; Aslett, González, Hamilton, & Pecho, 2024; Aslett, González, Hadwick, et al., 2024). En este estudio, el apoyo a la toma de decisiones se limita a ofrecer información descriptiva agregada; no se automatizan controles ni se producen clasificaciones individuales.

Los antecedentes aplicados muestran que los métodos no supervisados pueden reducir universos de revisión o identificar comportamientos atípicos (de Roux et al., 2018; Savić et al., 2022; Ordoñez & Hallo, 2023; Zheng et al., 2024). Sin embargo, esos trabajos emplean datos y objetivos distintos. La presente investigación no traslada sus conclusiones sobre fraude o riesgo: adopta únicamente la lección metodológica de combinar técnicas cuantitativas con explicación y cautela contextual.

Capítulo 3. Metodología

3.1 Enfoque y diseño de investigación

La investigación adopta un enfoque cuantitativo, con alcance descriptivo y exploratorio. Es cuantitativa porque trabaja con variables numéricas y categóricas derivadas de registros tributarios agregados. Es descriptiva porque busca caracterizar patrones presentes en los datos, y exploratoria porque aplica técnicas no supervisadas para identificar estructuras que no estaban definidas previamente.

El diseño es no experimental y longitudinal retrospectivo. No se manipulan variables ni se interviene sobre el fenómeno observado. Se analizan datos históricos del período 2017–2025, con el propósito de comprender su estructura agregada. Esta delimitación es coherente con el uso de datos abiertos y con el alcance académico del producto.

3.2 Referencia metodológica: CRISP-DM

El proceso se organiza tomando como referencia CRISP-DM, modelo que propone fases para estructurar proyectos de minería de datos desde la comprensión del problema hasta el despliegue de resultados (Wirth & Hipp, 2000). En este trabajo se utiliza como guía de ordenamiento, no como una receta rígida. También se relaciona con la lógica más amplia de descubrimiento de conocimiento en bases de datos, donde Fayyad et al. (1996) distinguen selección, preprocesamiento, transformación, minería de datos e interpretación.

La adaptación realizada fue la siguiente: la comprensión del negocio se refleja en el planteamiento del problema, objetivos y alcance; la comprensión de datos se desarrolló en el Notebook 01; la preparación se documentó en los Notebooks 01 y 02; el modelado y evaluación se trabajaron en el Notebook 03; y la comunicación de resultados se consolidó en el Notebook 04 y en la implementación del dashboard en Streamlit.

3.3 Relación entre CRISP-DM y el proyecto

Tabla 1

Relación entre CRISP-DM y el desarrollo del proyecto

Fase CRISP-DM	Aplicación en el proyecto	Evidencia principal
Comprensión del negocio	Problema, objetivos, alcance descriptivo y utilidad esperada para análisis fiscal.	Capítulo 1
Comprensión de los datos	Inventario de archivos, cobertura, estructura, compatibilidad y calidad inicial de las fuentes.	Notebook 01
Preparación de datos	Homologación, tratamiento de calidad, construcción de variables, escalamiento y selección de bases analíticas.	Notebooks 01 y 02
Modelado	Aplicación de K-Means, K-medoids, DBSCAN y agrupamiento jerárquico.	Notebook 03
Evaluación	Comparación mediante métricas internas, estabilidad, interpretabilidad y utilidad práctica.	Notebook 03
Despliegue	Preparación de tablas, gráficos, estructura conceptual e implementación del dashboard en Streamlit.	Notebook 04 y Notebook 05

Nota. Fuente: Wirth y Hipp (2000), adaptado a la estructura metodológica del proyecto. Elaboración: propia.

3.4 Fuentes de información y unidad de análisis

Las fuentes primarias corresponden a los archivos anuales de Recaudación y Ventas-Compras publicados por el SRI para 2017–2025. Se trabajó con dieciocho archivos y con sus estructuras disponibles. La integración se realizó únicamente sobre dimensiones compatibles, principalmente año, mes, provincia, cantón y sector económico.

La unidad de análisis es un registro agregado derivado de esas combinaciones. No corresponde a un contribuyente individual, una declaración completa ni una empresa. Esta definición evita atribuir los patrones del modelo a sujetos específicos y orienta la interpretación hacia relaciones temporales, territoriales, sectoriales y monetarias.

3.5 Preparación, modelado y evaluación

La preparación incluyó inventario, diagnóstico, homologación de nombres y tipos, revisión de faltantes y duplicados, tratamiento de ceros y valores extremos, construcción de

variables derivadas, selección de variables y escalamiento. Cada etapa produjo archivos intermedios y reportes de control que permiten reconstruir el flujo.

Se compararon K-Means, K-medoids, DBSCAN y agrupamiento jerárquico. K-Means se aplicó sobre la totalidad de las observaciones; los métodos de mayor costo se evaluaron sobre muestras reproducibles. La selección final combinó Silhouette, Davies-Bouldin y Calinski-Harabasz (Caliński & Harabasz, 1974) con el tamaño de clústeres, aplicabilidad, costo computacional y la interpretación. Como evaluación complementaria posterior a la selección, la solución final se contrastó con la clasificación institucional de grandes contribuyentes disponible para 2021–2025, sin incorporar esta variable al entrenamiento.

3.6 Implementación del dashboard y reproducibilidad

El dashboard se implementó en Streamlit como una aplicación multipágina. Sus módulos leen directamente los archivos CSV de resultados y no reentrenan el modelo. La separación entre procesamiento y visualización permite regenerar las tablas mediante los notebooks 01 a 04 y mantener la aplicación del Notebook 05 sin duplicar la fuente de datos.

La reproducibilidad se apoya en rutas documentadas, semillas aleatorias, archivos de verificación, inventarios y validaciones estructurales. Los gráficos y tablas del documento se derivan de los mismos resultados utilizados por el dashboard; por ello, las cifras deben coincidir entre notebooks, documento y aplicación.

3.7 Herramientas utilizadas

El desarrollo práctico se realizó principalmente en Python y Jupyter Notebook. pandas y NumPy se utilizaron para preparación y análisis; scikit-learn y librerías compatibles se emplearon para clustering y evaluación; Matplotlib y Plotly apoyaron la visualización; y

Streamlit permitió construir el dashboard. Los archivos parquet y CSV conservaron la separación entre datos intermedios, resultados y productos de consulta.

La implementación se mantuvo dentro del ecosistema de Python para reducir cambios de tecnología y conservar la trazabilidad. El dashboard funciona como prototipo académico: facilita la exploración de resultados, pero no reemplaza la revisión metodológica ni constituye una plataforma institucional en producción.

Capítulo 4. Desarrollo del proceso analítico integral aplicado a los datos abiertos tributarios del Ecuador correspondientes al período 2017–2025

4.1 Extracción de datos de Recaudación y Ventas-Compras

La primera etapa del desarrollo práctico se orientó a identificar, organizar y examinar las fuentes tributarias utilizadas en la investigación. Este proceso permitió verificar la cobertura temporal, la cantidad de archivos disponibles, la estructura de cada conjunto de datos y las dimensiones que podían emplearse para integrar la información sin alterar su nivel de agregación. La Tabla 2 sintetiza los principales resultados obtenidos durante esta fase.

Tabla 2

Resumen de extracción e integración inicial

Concepto	Resultado
Archivos considerados	18 archivos: 9 de Recaudación y 9 de Ventas-Compras
Período cubierto	2017–2025
Registros consolidados de Recaudación	5.893.618
Registros consolidados de Ventas-Compras	445.343
Dimensiones comunes utilizadas	Año, mes, provincia, cantón y sector económico
Base integrada preliminar	Construida únicamente sobre dimensiones compatibles

Nota. Fuente: Servicio de Rentas Internas (SRI), datos abiertos de Recaudación y Ventas-Compras, 2017–2025. Elaboración: propia.

Como se resume en la Tabla 2, el Notebook 01 permitió localizar, organizar e inventariar dieciocho archivos: nueve de Recaudación y nueve de Ventas-Compras, correspondientes a 2017–2025. Antes de la integración se consolidaron 5.893.618 registros de Recaudación y 445.343 de Ventas-Compras. Estas cantidades describen las fuentes originales consolidadas y no la dimensión final de la base de modelado.

El diagnóstico mostró que la integración debía realizarse sobre dimensiones compatibles y no sobre todas las variables disponibles. Por ello, se construyó una base integrada preliminar a

partir de año, mes, provincia, cantón y sector económico. Esta decisión permitió conservar una lectura agregada sin duplicar montos ni mezclar niveles de detalle incompatibles.

La salida principal de esta etapa fue la base base_integrada_preliminar.parquet, almacenada en la carpeta de datos intermedios junto con las bases agregadas por fuente. De esta manera, el Notebook 02 recibió un insumo reproducible y trazable para la preparación de variables.

4.2 Preparación de los datos tributarios obtenidos

Después de consolidar e integrar las fuentes compatibles, el Notebook 02 se orientó a preparar la información para el modelado. En esta etapa se revisaron la calidad y la estructura de los datos, se trataron valores faltantes, duplicados, categorías, ceros estructurales y valores extremos, y se construyeron variables derivadas. También se generaron matrices escaladas y versiones reducidas que permitieron comparar distintas alternativas de representación. La Tabla 3 resume las principales bases analíticas obtenidas como resultado de este proceso.

Tabla 3

Bases analíticas preparadas

Base generada	Dimensión	Uso
base_modelado_tratada.parquet	391.774 registros y 43 columnas	Base depurada con variables originales y derivadas
base_clustering_completa_escalada.parquet	391.774 registros y 27 columnas	Identificadores y 17 variables escaladas
base_clustering_reducida_escalada.parquet	391.774 registros y 24 columnas	Identificadores y 14 variables escaladas
base_pca_completa.parquet	391.774 registros y 16 columnas	Componentes principales para exploración
base_pca_reducida.parquet	391.774 registros y 16 columnas	Componentes principales de la base reducida

Nota. Fuente: Resultados del Notebook 02 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La base tratada quedó conformada por 391.774 registros y 43 columnas después de la revisión de faltantes, duplicados, categorías, ceros estructurales, valores extremos y diferencias

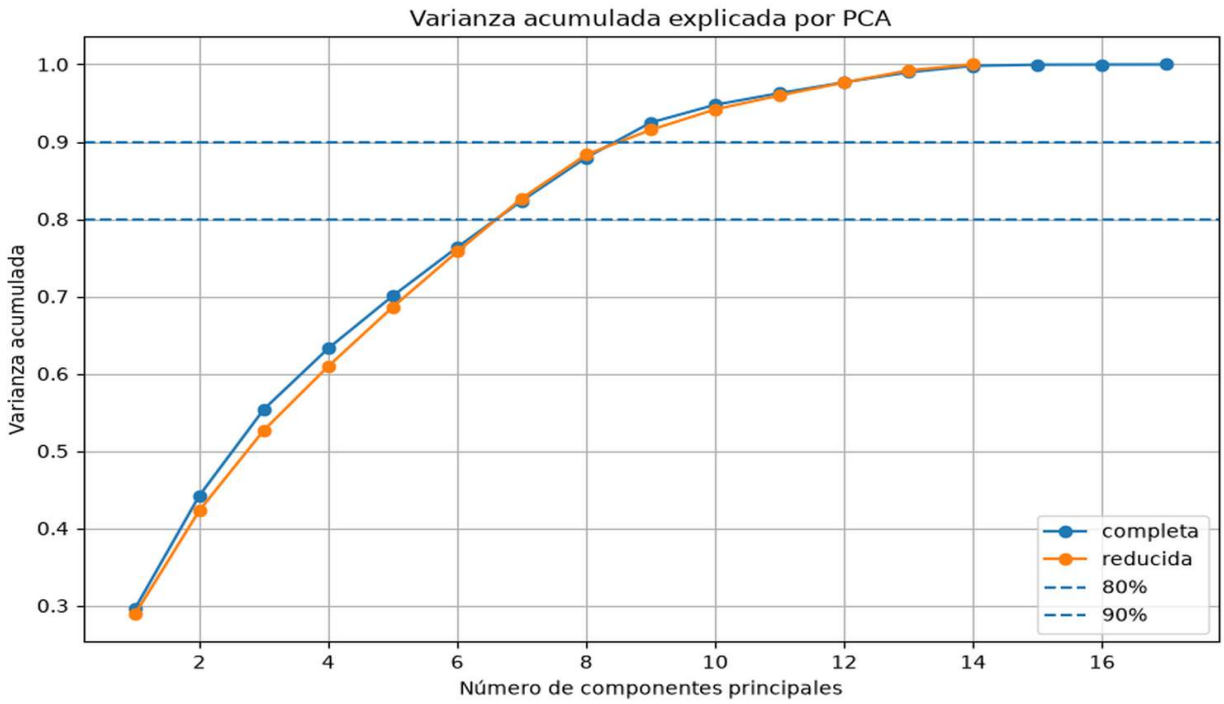
de escala. Además, se construyeron razones, indicadores relativos y transformaciones destinadas a representar relaciones entre las variables sin perder la referencia de sus valores originales.

Para el modelado se generaron matrices escaladas completa y reducida. La primera incluyó 17 variables y la segunda 14, después de excluir variables redundantes o menos útiles para la interpretación. La Figura 1 muestra que las versiones PCA presentan una acumulación de varianza similar; estos componentes se emplean como apoyo exploratorio y no reemplazan el análisis de las variables tributarias y económicas originales.

El análisis de componentes principales se utilizó como herramienta exploratoria complementaria. Su propósito fue revisar cuánto de la variabilidad podía resumirse en un número menor de componentes y disponer de una representación bidimensional para visualizar posteriormente los clústeres.

Figura 1

Varianza acumulada explicada por PCA en las versiones completa y reducida



Nota. Fuente: Resultados del Notebook 02 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

Las curvas de varianza acumulada de las versiones completa y reducida son muy similares. En ambos casos, el umbral del 80 % se supera alrededor del séptimo componente y el 90 % alrededor del noveno. Este resultado respalda que la versión reducida conserva una estructura informativa comparable, por lo que ambas versiones se mantuvieron como alternativas para la evaluación de modelos.

4.3 Modelado de patrones tributarios mediante técnicas de aprendizaje automático no supervisado

Una vez preparadas las matrices escaladas completa y reducida, el Notebook 03 desarrolló la etapa de modelado mediante cuatro técnicas de aprendizaje automático no supervisado: K-Means, K-medoids, DBSCAN y agrupamiento jerárquico. La aplicación de estos métodos permitió generar distintas soluciones candidatas y examinar la estructura de los datos desde enfoques basados en centroides, medoides, densidad y relaciones jerárquicas.

La implementación consideró las características y el volumen de la base analítica. K-Means pudo aplicarse sobre la totalidad de los registros, mientras que K-medoids y el agrupamiento jerárquico se ejecutaron sobre muestras reproducibles debido a su mayor costo computacional. DBSCAN se utilizó para explorar la formación de grupos por densidad y la presencia de observaciones clasificadas como ruido. Estas diferencias de aplicación se mantuvieron documentadas para evitar que los resultados fueran interpretados como si todos los modelos hubieran sido ejecutados bajo las mismas condiciones.

Para K-Means se generaron configuraciones con diferentes cantidades de clústeres sobre las versiones completa y reducida de la matriz de modelado. Cada configuración produjo una asignación de grupos y los resultados necesarios para su posterior evaluación. Las bases construidas mediante componentes principales se utilizaron únicamente como apoyo para

representar visualmente las soluciones, sin sustituir las variables tributarias y económicas originales.

Después de generar las soluciones candidatas, se procedió a compararlas mediante métricas internas, aplicabilidad sobre la base completa, costo computacional e interpretabilidad. Los resultados de esta evaluación se presentan en el apartado siguiente.

4.4 Evaluación de los modelos de agrupamiento

La evaluación se desarrolló en tres momentos. En primer lugar, se compararon las soluciones candidatas mediante métricas internas y criterios prácticos. En segundo lugar, se examinó con mayor detalle la solución seleccionada para identificar el tamaño, la composición y las características principales de sus clústeres. En tercer lugar, se realizó un contraste externo de la solución seleccionada con la clasificación institucional de grandes contribuyentes disponible en los archivos de Recaudación del SRI.

4.4.1 Comparación y selección del modelo

En la primera fase se compararon las soluciones obtenidas mediante K-Means, K-medoids, DBSCAN y agrupamiento jerárquico con el método de Ward. La Tabla 4 presenta los resultados generales de esta comparación y distingue entre los modelos aplicados a toda la base y aquellos evaluados sobre muestras reproducibles.

Tabla 4

Resumen comparativo de modelos candidatos

Modelo candidato	Silhouette	Davies-Bouldin	Comentario de evaluación
K-Means reducida, k = 3	0,3624	1,2579	Solución interpretable, escalable y aplicada a toda la base
K-Means completa, k = 2	0,2196	1,8038	Solución más simple, pero menos diferenciada
K-Medoids reducida, k = 2	0,4098	1,2587	Buena métrica, evaluada sobre muestra reproducible
Jerárquico Ward, k = 3	0,4404	1,0799	Buen desempeño muestral; menor escalabilidad

Modelo candidato	Silhouette	Davies-Bouldin	Comentario de evaluación
DBSCAN sobre la base reducida	No comparable	No comparable	Útil para analizar ruido y valores atípicos; no seleccionado como modelo principal

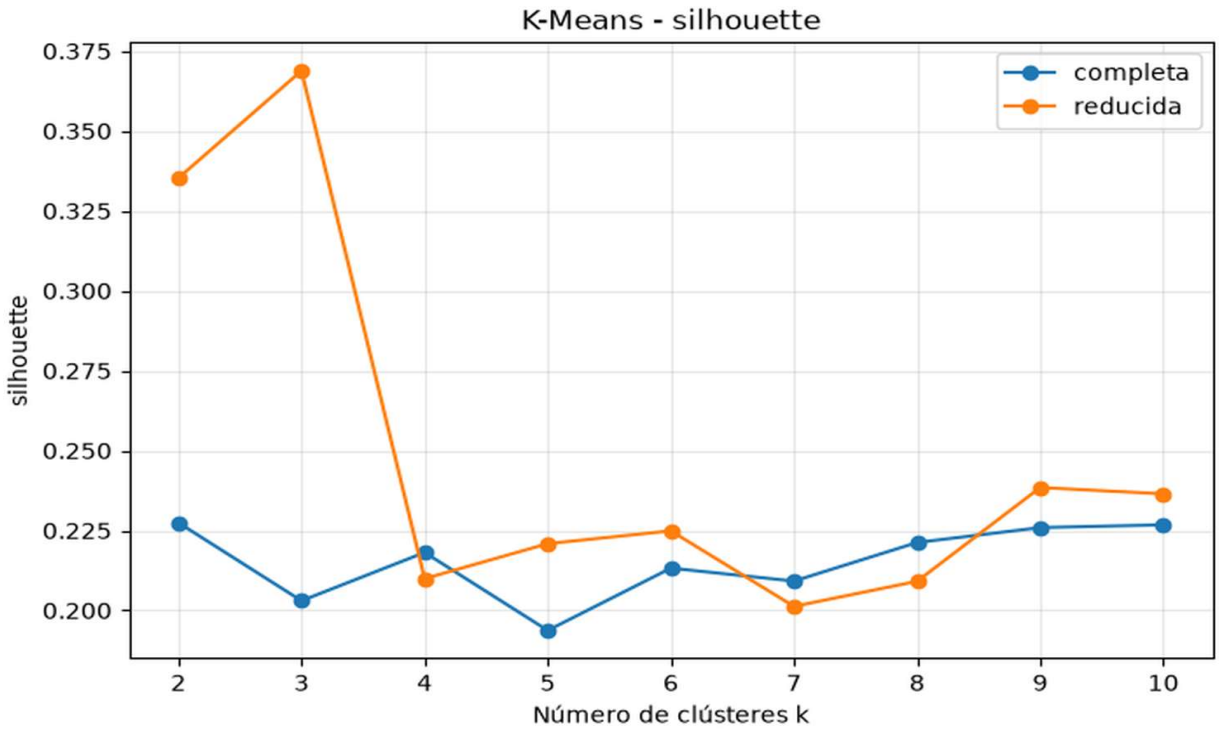
Nota. Fuente: Resultados de evaluación del Notebook 03 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La Tabla 4 distingue los resultados obtenidos sobre la base completa de aquellos calculados mediante muestras reproducibles. Esta diferencia es metodológicamente importante, porque las métricas muestrales permiten contrastar los enfoques, pero no deben interpretarse como si todos los modelos hubieran sido aplicados bajo las mismas condiciones.

Entre las configuraciones de K-Means, la versión reducida con tres clústeres presentó el mejor coeficiente Silhouette. La Figura 2 muestra el comportamiento de esta métrica al variar el número de clústeres en las matrices completa y reducida.

Figura 2

Comparación del coeficiente Silhouette de K-Means según número de clústeres



Nota. Fuente: Resultados preliminares del Notebook 03 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

En la evaluación preliminar, la versión reducida con k = 3 presentó el mayor coeficiente Silhouette entre las configuraciones de K-Means consideradas. En la muestra utilizada para la comparación final, esta solución obtuvo un valor de 0,3624. La diferencia entre ambas mediciones se relaciona con el momento y la muestra empleados en cada evaluación; sin embargo, en los dos casos, k = 3 se mantuvo como la alternativa más destacada entre las configuraciones de K-Means.

Las métricas se interpretaron de manera conjunta. La inercia disminuyó a medida que aumentó el número de clústeres, mientras que Silhouette, Davies-Bouldin y Calinski-Harabasz favorecieron configuraciones diferentes. Por esta razón, la selección no se sustentó en una sola cifra, sino en la combinación del desempeño técnico, el tamaño de los grupos, la aplicabilidad sobre la base completa, el costo computacional y la interpretación tributaria y económica.

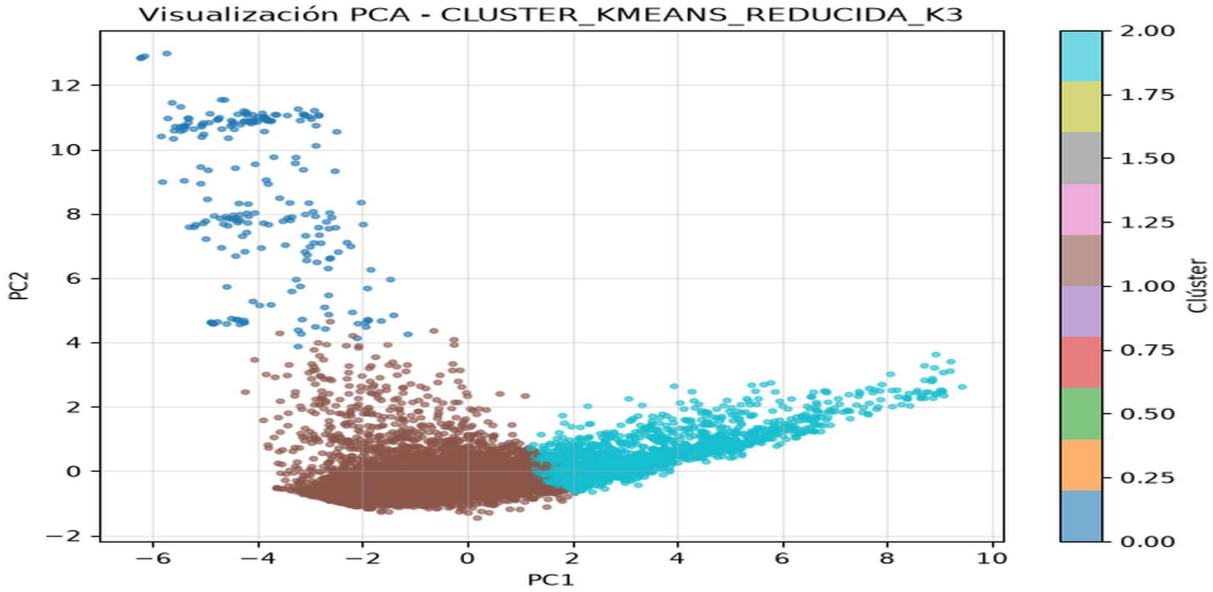
Aunque K-medoids y el agrupamiento jerárquico con Ward presentaron valores favorables en las muestras utilizadas, su mayor costo computacional limitó su aplicación sobre la totalidad de los registros. La solución K-Means sobre la base reducida con tres clústeres ofreció un equilibrio más adecuado entre separación, escalabilidad e interpretabilidad. Por ello, fue seleccionada como modelo final.

4.4.2 Resultados del modelo seleccionado

Una vez seleccionada la solución K-Means reducida con tres clústeres, se examinó su distribución en el plano de componentes principales y se analizaron el tamaño, la composición y las características de los grupos obtenidos. La Figura 3 presenta una representación bidimensional de la solución final.

Figura 3

Distribución de la solución K-Means reducida con tres clústeres en el plano PCA



Nota. Fuente: Resultados del Notebook 03; muestra gráfica reproducible de 10.000 observaciones derivadas de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La representación sobre PC1 y PC2 distingue una estructura simple y comprensible: un grupo asociado a valores altos de PC2, un grupo central de alta concentración y otro más extendido hacia valores positivos de PC1. La visualización funciona como apoyo interpretativo y no como evidencia única de calidad, ya que los componentes representan una reducción bidimensional de un espacio de mayor dimensión.

Tabla 5

Perfil resumido de clústeres

Clúster	Nombre interpretativo	Registros	Porcentaje	Característica principal
0	Ventas bajas y compras relativamente altas	7.609	1,94 %	Grupo pequeño con ventas bajas, compras relativamente altas y razones elevadas
1	Actividad frecuente de menor magnitud	316.483	80,78 %	Grupo mayoritario con valores monetarios bajos o medios
2	Alta magnitud económica y comercio exterior	67.682	17,28 %	Grupo con mayor recaudación, ventas, compras y comercio exterior

Nota. Fuente: Resultados del modelo K-Means reducido k = 3 obtenidos en el Notebook 03 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La Tabla 5 resume los tres perfiles. El clúster 0, ventas bajas y compras relativamente altas, reunió 7.609 registros (1,94 %); el clúster 1, actividad frecuente de menor magnitud, concentró 316.483 (80,78 %); y el clúster 2, alta magnitud económica y comercio exterior, agrupó 67.682 (17,28 %). Estas denominaciones son etiquetas interpretativas construidas a partir de medianas y variables diferenciadoras, no categorías oficiales del SRI.

La Tabla 6 y la Figura 4 muestran la diferencia entre frecuencia y magnitud: el clúster de actividad frecuente reúne la mayoría de observaciones, mientras que el de alta magnitud concentra el 96,51 % de la recaudación. La Figura 5 confirma que los tres perfiles aparecen durante todo el período, aunque su participación cambia gradualmente. Este patrón describe composición agregada y no seguimiento longitudinal de contribuyentes.

Tabla 6

Participación de registros y concentración monetaria por clúster

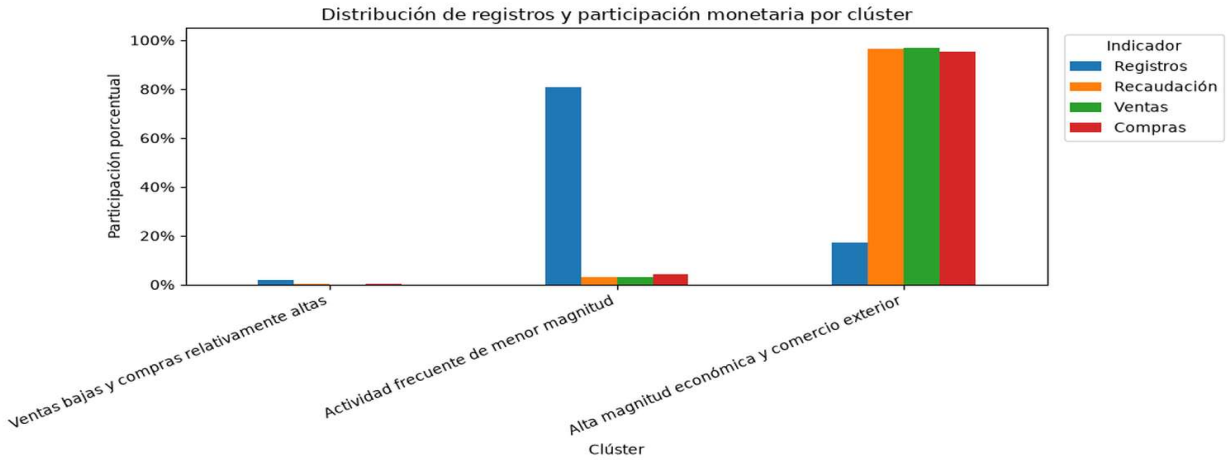
Clúster	Registros	% registros	% recaudación	% ventas	% compras
Ventas bajas y compras relativamente altas	7.609	1,94 %	0,48 %	0,00 %	0,46 %
Actividad frecuente de menor magnitud	316.483	80,78 %	3,02 %	3,16 %	4,36 %
Alta magnitud económica y comercio exterior	67.682	17,28 %	96,51 %	96,84 %	95,18 %

Nota. Fuente: Cálculos sobre la base final de clústeres generada en el Notebook 03 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

El clúster Actividad frecuente de menor magnitud representa el 80,78 % de los registros, pero solo el 3,02 % de la recaudación. En contraste, Alta magnitud económica y comercio exterior reúne el 17,28 % de los registros y concentra el 96,51 % de la recaudación. El clúster de ventas bajas y compras relativamente altas mantiene una participación reducida tanto en registros como en montos agregados.

Figura 4

Distribución de registros y participación monetaria por clúster

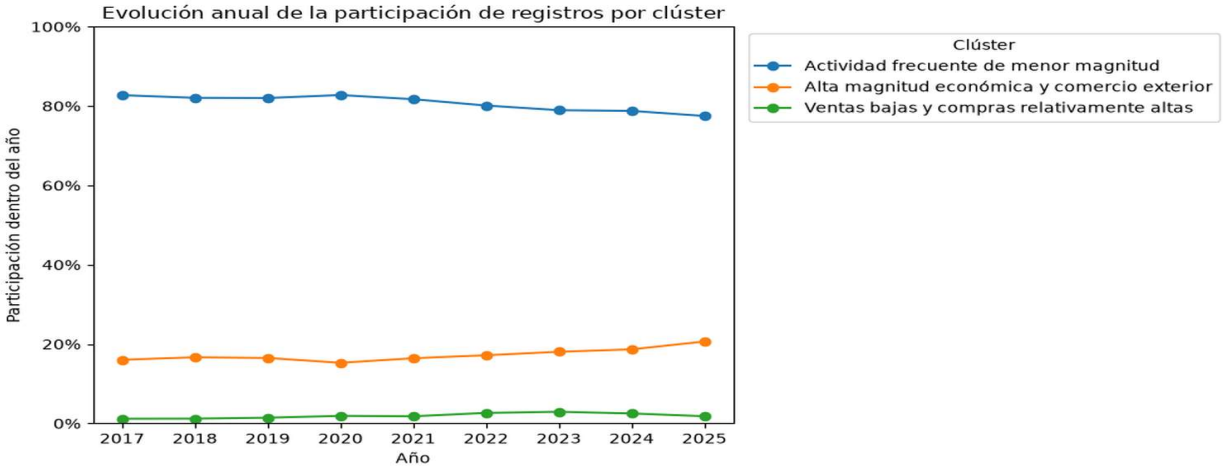


Nota. Fuente: Resultados consolidados en el Notebook 04 a partir de datos abiertos del SRI, 2017-2025. Elaboración: propia.

La figura sintetiza la principal diferencia estructural del modelo: la frecuencia de observaciones y la magnitud económica no se distribuyen de la misma forma. El grupo mayoritario en registros presenta una participación monetaria limitada, mientras que el grupo de alta magnitud domina los valores de recaudación, ventas y compras.

Figura 5

Evolución anual de la participación de registros por clúster



Nota. Fuente: Resultados temporales del Notebook 04 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

Los tres clústeres se observan durante todo el período, lo que aporta estabilidad interpretativa. La participación de Actividad frecuente de menor magnitud disminuyó de 82,75 % en 2017 a 77,51 % en 2025, mientras que Alta magnitud económica y comercio exterior aumentó de 16,05 % a 20,66 %. El clúster de ventas bajas y compras relativamente altas se mantuvo como el grupo más pequeño y alcanzó su mayor participación en 2023, con 2,96 %.

4.4.3 Validación externa institucional de la solución seleccionada

Como complemento a las métricas internas utilizadas para evaluar los modelos, se realizó un contraste externo de la solución K-Means seleccionada con la clasificación de grandes contribuyentes disponible en los archivos originales de Recaudación del SRI. Esta variable no intervino en el entrenamiento del modelo ni en la construcción de las variables utilizadas para el agrupamiento, por lo que se empleó únicamente como criterio institucional posterior de comparación. El análisis se limitó al período 2021–2025, correspondiente a la vigencia de esta segmentación en las fuentes utilizadas.

Para mantener la misma unidad de análisis del modelo, la recaudación asociada a grandes contribuyentes se agregó mediante la clave compuesta por año, mes, provincia, cantón y sector económico. Posteriormente, estos montos se integraron con las etiquetas de los clústeres. El procedimiento conservó los 214.389 registros de la base correspondientes a 2021–2025, sin observaciones sin correspondencia. La recaudación reconstruida desde la fuente original coincidió con los valores de la base analítica; las diferencias residuales fueron inferiores a $0,000001 y correspondieron a precisión numérica.

El contraste mostró una correspondencia marcada con el perfil Alta magnitud económica y comercio exterior. De los $43.099,95 millones de recaudación asociados a grandes contribuyentes durante 2021–2025, el 99,20 % se concentró en el clúster 2. Esta relación se

mantuvo en cada año del período de validación: 99,63 % en 2021; 99,44 % en 2022; 97,68 % en 2023; 99,56 % en 2024; y 99,66 % en 2025. En contraste, los clústeres 0 y 1 reunieron 0,47 % y 0,33 %, respectivamente, de la recaudación acumulada de este segmento.

La convergencia observada es consistente con los criterios institucionales establecidos para la consideración de grandes contribuyentes, entre los que se incluyen la relevancia de las transacciones, el aporte a la recaudación y el comportamiento del sector económico (Servicio de Rentas Internas, 2021). Sin embargo, este resultado no implica equivalencia entre ambas clasificaciones. La categoría institucional corresponde a sujetos pasivos, mientras que los clústeres representan combinaciones agregadas de período, territorio y actividad económica. En consecuencia, el contraste aporta evidencia externa sobre la coherencia descriptiva del perfil de alta magnitud, pero no permite identificar ni clasificar contribuyentes individuales.

4.5 Implementación de un dashboard interactivo

Los Notebooks 04 y 05 permitieron consolidar los resultados del proceso e implementar el dashboard interactivo en Streamlit. La aplicación consume directamente las tablas finales generadas durante las etapas anteriores, sin recalcular las etiquetas del modelo, y organiza su contenido en cuatro páginas orientadas a tareas analíticas diferentes. La Tabla 7 resume la estructura funcional adoptada para presentar los indicadores y facilitar la exploración de los resultados.

Tabla 7

Estructura funcional final del dashboard

Página	Objetivo analítico	Componentes principales	Insumos principales
Resumen ejecutivo	Presentar el panorama general y los principales hallazgos	Indicadores, evolución anual, contraste de clústeres y concentración provincial	Temporal, provincia, síntesis y dimensión de clústeres
Evolución temporal	Examinar cambios anuales y composición por clúster	Filtros, tendencia, variación interanual, composición y mapa de calor	dashboard_temporal_clusters.csv

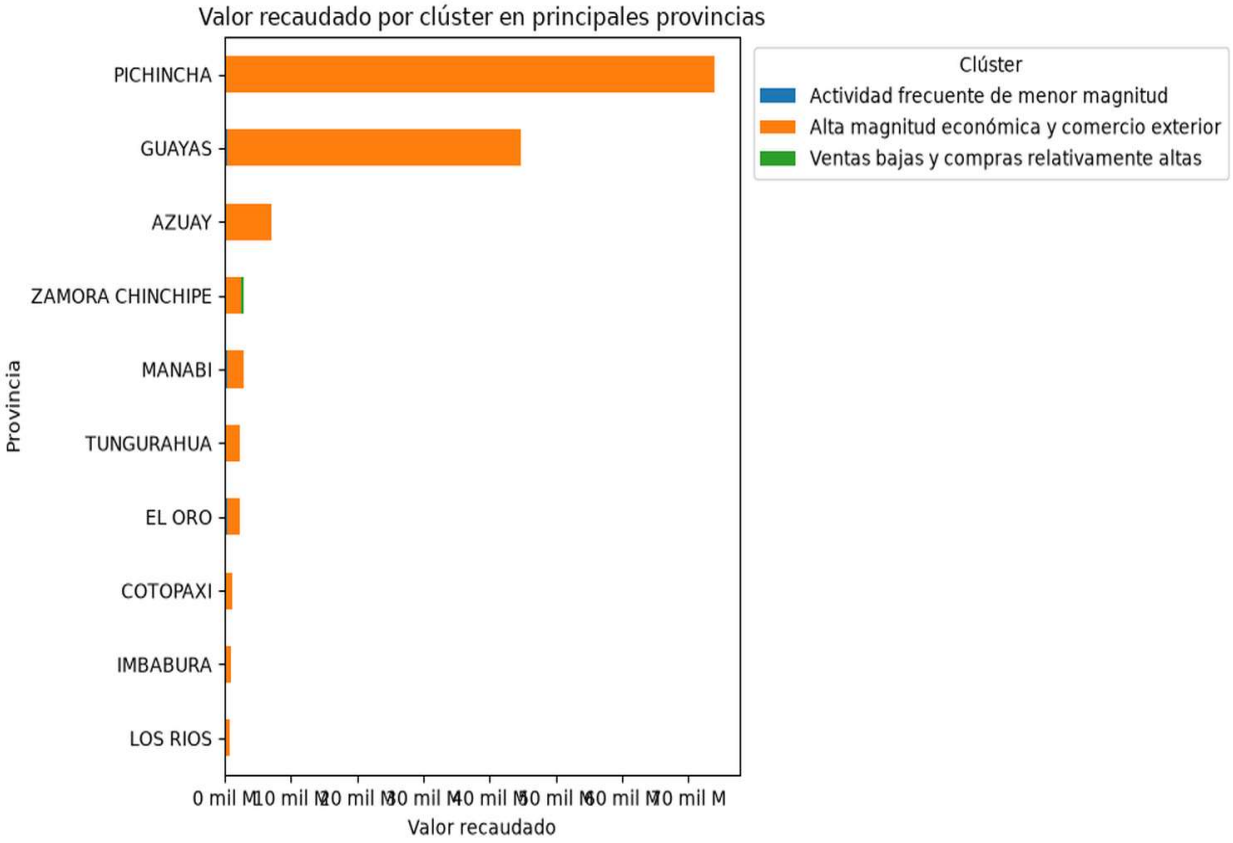
Página	Objetivo analítico	Componentes principales	Insumos principales
Análisis territorial	Comparar provincias y profundizar a nivel cantonal	Ranking provincial, composición, concentración y detalle cantonal	dashboard_provincia_clusters.csv; dashboard_canton_clusters.csv
Perfiles económicos	Analizar relaciones entre ventas, compras, recaudación y sectores	Dispersión, medianas por clúster, sectores y composición sectorial	dashboard_base_clusters.csv; tablas sectoriales con nombres

Nota. Fuente: Estructura de la aplicación desarrollada en el Notebook 05 y tablas preparadas por los Notebooks 03 y 04. Elaboración: propia.

Las Figuras 6 y 7 ejemplifican las vistas territoriales y sectoriales construidas a partir de las tablas finales. Cada página utiliza filtros, indicadores y gráficos complementarios para responder a una tarea analítica específica, sin modificar los resultados ni recalcular el modelo dentro de Streamlit.

Figura 6

Valor recaudado por clúster en las principales provincias

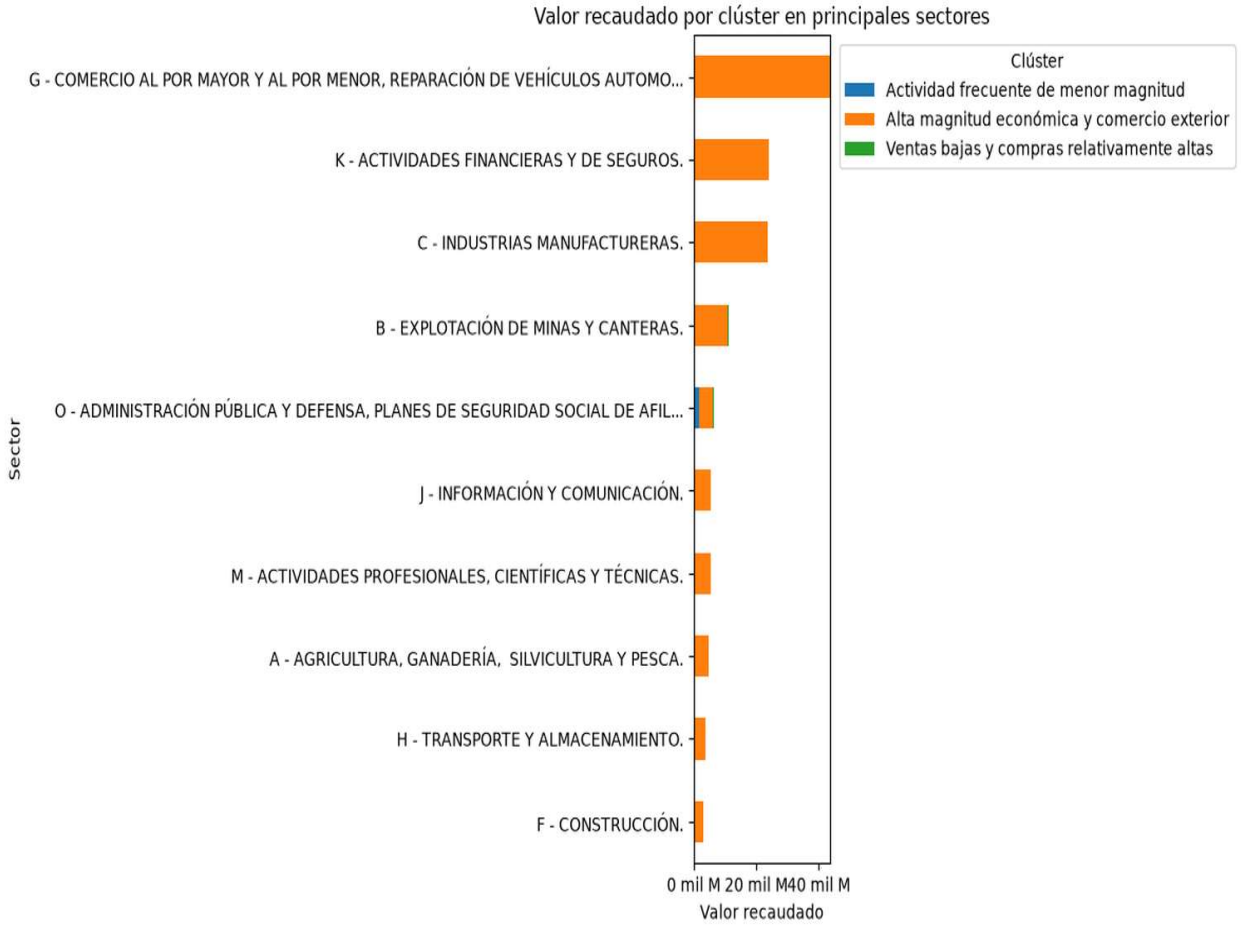


Nota. Fuente: Resultados territoriales del Notebook 04 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

Pichincha y Guayas concentran los mayores montos de recaudación y superan ampliamente al resto de provincias. En las principales provincias, el clúster Alta magnitud económica y comercio exterior explica la mayor parte del valor recaudado. Esta lectura confirma que el orden territorial por número de registros no coincide necesariamente con el orden por magnitud monetaria.

Figura 7

Valor recaudado por clúster en los principales sectores económicos



Nota. Fuente: Resultados sectoriales del Notebook 04 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

El comercio al por mayor y al por menor registra la mayor recaudación sectorial, seguido por actividades financieras y de seguros, industrias manufactureras y explotación de minas y canteras. En los sectores líderes, la mayor parte de la recaudación corresponde al clúster de alta magnitud económica, lo que evidencia que la concentración monetaria tiene simultáneamente una dimensión territorial y sectorial.

El dashboard permite sintetizar los resultados del proceso y explorarlos mediante filtros de año, clúster, territorio y sector. Los indicadores que se presentan a continuación corresponden a las tablas oficiales generadas por los notebooks; el dashboard organiza su consulta, pero no modifica los resultados del modelo.

Tabla 8

Indicadores principales del dashboard para 2025

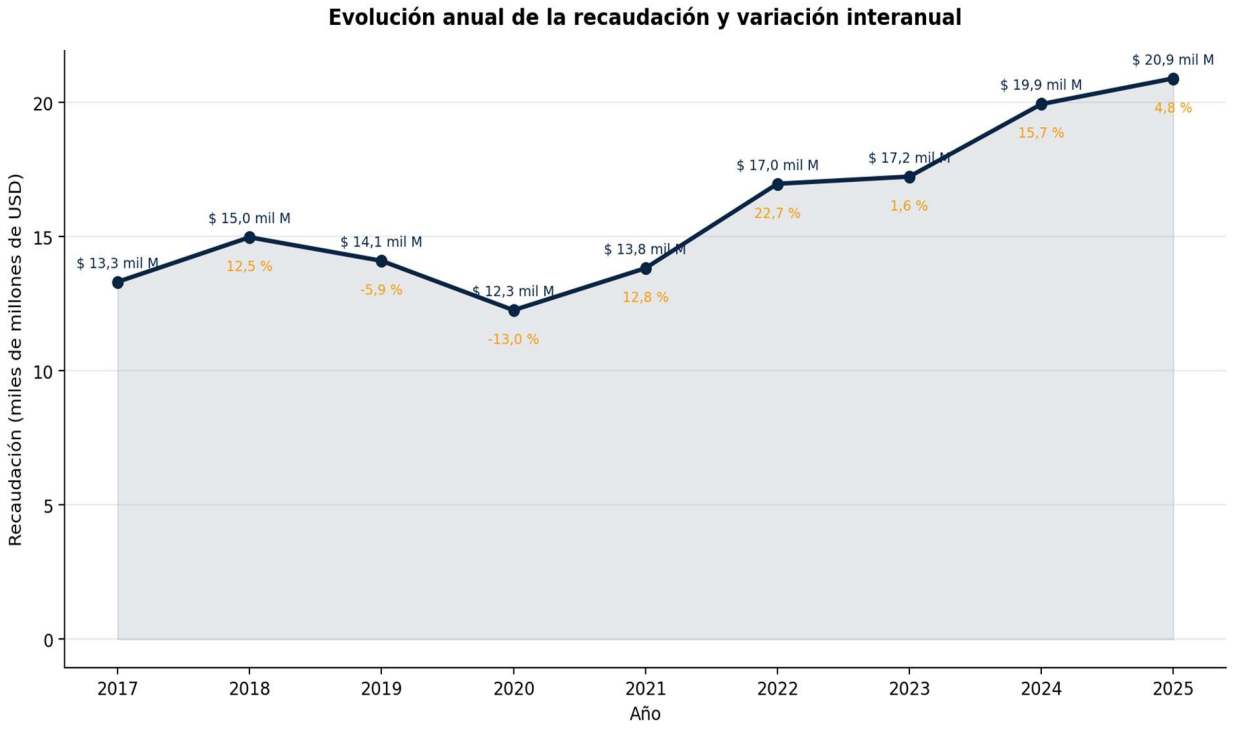
Indicador	Valor 2025	Variación frente a 2024	Lectura
Registros agregados	43.302	-0,58 %	El volumen de observaciones se mantuvo relativamente estable
Recaudación total	$ 20.900.488.321,68	4,81 %	La recaudación alcanzó el mayor nivel del período
Recaudación por registro	$ 482.667,97	5,42 %	El monto promedio por observación agregada continuó aumentando

Nota. Fuente: Dashboard Streamlit; cálculos sobre dashboard_temporal_clusters.csv generado a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La Tabla 8 resume los indicadores de 2025: 43.302 observaciones agregadas y una recaudación de $ 20.900.488.321,68. Frente a 2024, el monto aumentó 4,81 % y el número de registros disminuyó 0,58 %; por ello, la recaudación media por registro alcanzó $ 482.667,97 y creció 5,42 %. La Figura 8 sitúa este cambio dentro de la serie anual completa.

Figura 8

Evolución anual de la recaudación y variación interanual

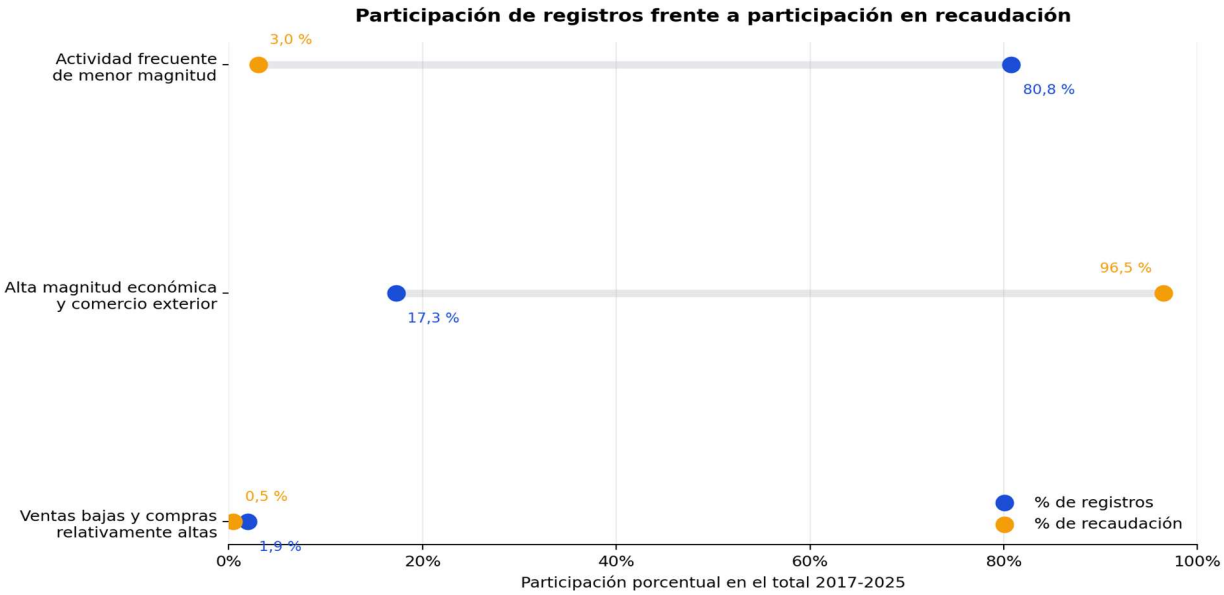


Nota. Fuente: Dashboard Streamlit; cálculos sobre tablas temporales generadas a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La recaudación pasó de aproximadamente $ 13,3 mil millones en 2017 a $ 20,9 mil millones en 2025. El período presenta reducciones en 2019 y 2020, seguidas por una recuperación desde 2021. El mayor crecimiento interanual se observó en 2022, con 22,75 %, y en 2024 se registró otro aumento importante de 15,67 %. En 2025 el crecimiento fue positivo, aunque más moderado, con 4,81 %.

Figura 9

Participación de registros frente a participación en recaudación por clúster



Nota. Fuente: Dashboard Streamlit; cálculos sobre dashboard_base_clusters.csv generado a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La Figura 9 compara directamente la participación de registros y de recaudación. Actividad frecuente de menor magnitud reúne 80,78 % de las observaciones y 3,02 % del monto; Alta magnitud económica y comercio exterior representa 17,28 % y concentra 96,51 %, respectivamente. La diferencia confirma que la segmentación está dominada tanto por escala como por frecuencia.

Tabla 9

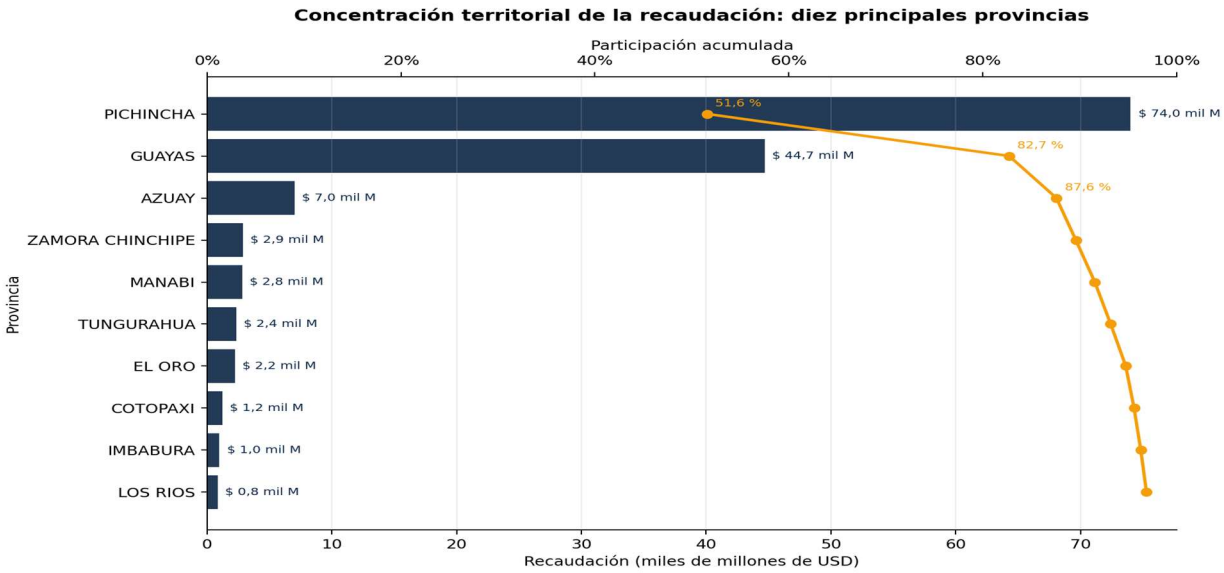
Principales provincias por recaudación acumulada

Provincia	Registros	Recaudación	Participación	Acumulada
Pichincha	16.815	$ 74.025,3 millones	51,57 %	51,57 %
Guayas	46.224	$ 44.702,5 millones	31,14 %	82,71 %
Azuay	24.519	$ 7.003,9 millones	4,88 %	87,59 %
Zamora Chinchipe	14.983	$ 2.873,6 millones	2,00 %	89,59 %
Manabí	40.535	$ 2.806,9 millones	1,96 %	91,55 %

Nota. Fuente: Dashboard Streamlit; cálculos sobre dashboard_provincia_clusters.csv generado a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

Figura 10

Concentración territorial de la recaudación en las diez principales provincias



Nota. Fuente: Dashboard Streamlit; cálculos sobre tablas provinciales generadas a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La Tabla 9 y la Figura 10 muestran una concentración territorial elevada. Pichincha reúne el 51,57 % y Guayas el 31,14 % de la recaudación; juntas acumulan el 82,71 %. Al incorporar Azuay, Zamora Chinchipe y Manabí, las cinco primeras provincias alcanzan el 91,55 %. Esta concentración describe la distribución de los montos en la base y no permite atribuir por sí sola productividad, cumplimiento o capacidad fiscal a cada territorio.

Tabla 10

Principales sectores por recaudación y recaudación por registro

Código	Sector	Registros	Recaudación total	Recaudación por registro
G	Comercio al por mayor y al por menor, reparación de vehículos automotores y motocicletas	23.609	$ 43.697,6 millones	$ 1.850.888,23
K	Actividades financieras y de seguros	11.350	$ 24.040,0 millones	$ 2.118.061,25
C	Industrias manufactureras	23.180	$ 23.595,4 millones	$ 1.017.919,96
B	Explotación de minas y canteras	10.686	$ 11.089,4 millones	$ 1.037.745,90
O	Administración pública y defensa, planes de seguridad social de afiliación obligatoria	22.274	$ 6.485,6 millones	$ 291.175,06

Nota. Fuente: Dashboard Streamlit; cálculos sobre dashboard_resumen_sectores_con_nombres.csv generado a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

La Tabla 10 presenta dos rankings que no deben confundirse. El sector G —comercio al por mayor y al por menor y reparación de vehículos— registra la mayor recaudación acumulada, con aproximadamente $ 43.697,6 millones; en cambio, el sector K —actividades financieras y de seguros— alcanza la mayor recaudación por registro, con $ 2.118.061,25. El primer indicador mide monto total y el segundo una razón agregada, no rentabilidad ni carga efectiva individual.

4.6 Análisis de resultados

Los resultados responden a las preguntas de investigación mediante una cadena verificable: los archivos se inventariaron, las estructuras se compararon, la integración se limitó a dimensiones compatibles, las variables se prepararon, los modelos se evaluaron y los hallazgos se comunicaron en el dashboard. Esta secuencia confirma la advertencia de Janssen et al. (2012): la disponibilidad de datos abiertos no equivale a reutilización inmediata. El valor analítico surgió del trabajo de comprensión, calidad y documentación.

La preparación fue decisiva porque la calidad no se redujo a eliminar nulos. La revisión de categorías, tipos, ceros, valores extremos y granularidad coincide con la concepción multidimensional de Wang y Strong (1996) y con el proceso de diagnóstico y mejora descrito por Batini et al. (2009). La decisión de no unir dimensiones incompatibles es un resultado metodológico en sí mismo: evita duplicaciones y una apariencia de precisión que las fuentes no sostienen.

La comparación de algoritmos también confirma que ninguna métrica debe decidir por sí sola. K-medoids y el método jerárquico obtuvieron resultados muestrales competitivos, pero su costo dificultó aplicarlos sobre 391.774 observaciones. K-Means reducida con tres clústeres ofreció un balance entre separación, escalabilidad e interpretación, consistente con la utilidad

práctica atribuida a K-Means (MacQueen, 1967) y con la recomendación de evaluar el clustering según datos y propósito (Xu & Wunsch, 2009).

La distribución de los clústeres muestra que frecuencia y magnitud económica son dimensiones diferentes. El grupo mayoritario concentra 80,78 % de los registros y solo 3,02 % de la recaudación, mientras que el grupo de alta magnitud reúne 17,28 % de los registros y 96,51 % del monto. Este contraste es el hallazgo central: una lectura basada únicamente en conteos ocultaría la concentración monetaria; una lectura basada solo en montos invisibilizaría la extensión del patrón frecuente.

El contraste externo con la clasificación institucional de grandes contribuyentes aporta evidencia complementaria sobre la coherencia del perfil de alta magnitud. Sin haber utilizado esta variable durante el entrenamiento, el clúster Alta magnitud económica y comercio exterior concentró el 99,20 % de la recaudación asociada a grandes contribuyentes durante 2021–2025. Esta correspondencia se mantuvo en cada año del período de validación, con participaciones entre 97,68 % y 99,66 %. La convergencia es consistente con los criterios institucionales de relevancia económica y aporte recaudatorio considerados por el SRI (Servicio de Rentas Internas, 2021), aunque no implica equivalencia entre ambas clasificaciones, debido a que el modelo agrupa observaciones agregadas y no contribuyentes individuales.

Aunque los tres clústeres fueron caracterizados a partir de sus perfiles estadísticos y variables diferenciadoras, el clúster 0 requiere una revisión complementaria debido a que corresponde al grupo de menor tamaño y presenta un comportamiento que requiere mayor cautela interpretativa. En particular, la combinación de ventas relativamente bajas, compras elevadas y razones altas de recaudación respecto de la actividad económica podría conducir a interpretaciones incorrectas si se analiza únicamente a partir de los centroides o de los montos

agregados. Por esta razón, se profundizó su composición sectorial, territorial y temporal utilizando las etiquetas ya generadas por K-Means, sin modificar el modelo ni reasignar observaciones.

El análisis confirmó inicialmente que la denominación del perfil representa un comportamiento predominante y no una condición absoluta: de los 7.609 registros asignados al clúster 0, 6.789 (89,22 %) presentaron compras superiores a ventas. La mediana de compras alcanzó $19.618,44, frente a una mediana de ventas de $52,50. Su composición sectorial tampoco fue homogénea. Administración pública y defensa concentró 3.053 registros, equivalentes al 40,12 % del clúster, constituyéndose en la actividad con mayor presencia dentro de este perfil. Sin embargo, el análisis territorial mostró que otros sectores pueden adquirir relevancia bajo contextos específicos, como ocurre con Explotación de minas y canteras en Zamora Chinchipe. En consecuencia, el patrón de compras superiores a ventas no corresponde a una única actividad económica, sino que puede manifestarse bajo estructuras sectoriales y territoriales diferentes.

La Tabla 11 resume los principales indicadores obtenidos en este análisis complementario.

Tabla 11

Indicadores complementarios para la interpretación del clúster 0

Dimensión	Indicador	Resultado
Perfil general	Registros del clúster 0	7.609 (1,94 %)
Relación compras–ventas	Registros con compras > ventas	6.789 (89,22 %)
Magnitud típica	Mediana de ventas / compras	$52,50 / $19.618,44
Sectorial	Administración pública y defensa dentro del clúster 0	3.053 (40,12 %)
Territorial-sectorial	Minería en Zamora Chinchipe	61 registros
Territorial-sectorial	Participación de minería en la recaudación del clúster 0 de Zamora Chinchipe	98,09 %
Minería–Zamora	Ventas / compras / recaudación	$3,30 / $499,55 / $325,19 millones
Tributaria	Principales componentes de los $325,19 millones	Renta sociedades: 31,63 %; regalías anticipadas: 21,53 %; utilidades mineras: 20,47 %

Dimensión	Indicador	Resultado
Temporal particular	El Pangui, minería, 2022	$1.057,36 millones en ventas; $469,83 millones en compras; ventas > compras en 9 de 12 meses
Temporal general	Participación mensual media del clúster 0	Enero–mayo: 2,24 %; junio–diciembre: 1,76 %

Nota. Los indicadores generales, sectoriales y temporales se calcularon sobre la base final de clústeres generada a partir de datos abiertos del SRI para 2017–2025. La composición tributaria de Explotación de minas y canteras en Zamora Chinchipe se obtuvo recuperando el detalle de los archivos originales de Recaudación mediante la misma clave de integración utilizada en el proceso analítico: año, mes, provincia, cantón y sector. Elaboración: propia.

Esta heterogeneidad se observa con especial claridad en Zamora Chinchipe, donde Explotación de minas y canteras reunió 61 registros del clúster 0 y concentró el 98,09 % de la recaudación correspondiente a dicho clúster en la provincia. En estas observaciones, el 95,08 % presentó compras superiores a ventas y, en conjunto, registraron aproximadamente $499,55 millones en compras, $3,30 millones en ventas y $325,19 millones de recaudación. La recuperación del detalle de la fuente original de Recaudación permitió desagregar la composición de este monto, en la que predominaron Renta sociedades (31,63 %), regalías anticipadas (21,53 %) y utilidades de las actividades mineras (20,47 %), mientras que las retenciones en la fuente representaron 5,70 %.

La dimensión temporal permite precisar esta aparente contradicción entre elevada recaudación y bajas ventas corrientes. En 2023, únicamente dos observaciones mineras de Zamora Chinchipe fueron asignadas al clúster 0; estas acumularon $172,18 millones de recaudación, frente a $0,21 millones en ventas y $34,70 millones en compras. Dentro de esa recaudación, Renta sociedades representó 59,68 % y Utilidades de las actividades mineras 38,61 %. Sin embargo, la integración de las fuentes se realiza a partir de combinaciones de año, mes, provincia, cantón y sector, por lo que la recaudación registrada en una determinada combinación no demuestra que la obligación tributaria se haya originado en las ventas y compras registradas en ese mismo mes. Esta consideración resulta especialmente relevante porque, en El Pangui durante 2022, el sector minero registró $1.057,36 millones en ventas y $469,83 millones en

compras, con ventas superiores a compras en 9 de los 12 meses. El resultado es compatible con una diferencia temporal entre la actividad económica observada y determinados flujos de recaudación posteriores, aunque el nivel agregado de los datos no permite vincular contribuyentes ni establecer una relación causal entre ambos períodos.

Finalmente, el clúster 0 también mostró un componente temporal recurrente a escala general. Su participación mensual media fue aproximadamente 2,24 % entre enero y mayo, frente a 1,76 % entre junio y diciembre, y en cada uno de los nueve años analizados, el máximo mensual se ubicó entre enero y abril. El patrón no fue uniforme, pues su intensidad varió entre años, por lo que debe interpretarse como un comportamiento temporal descriptivo de mayor presencia relativa durante los primeros meses y no como evidencia de estacionalidad formal. En conjunto, los resultados indican que el clúster 0 combina componentes sectoriales, territoriales y temporales y que sus razones elevadas pueden estar condicionadas por la granularidad y temporalidad de las fuentes. Aunque de Roux et al. (2018), Savić et al. (2022) y Ordoñez y Hallo (2023) han empleado métodos no supervisados en contextos relacionados con riesgo o anomalías tributarias, sus objetivos y datos difieren de los utilizados en este trabajo. Por ello, el perfil identificado debe entenderse como una descripción de observaciones agregadas y no como evidencia de evasión, incumplimiento o una situación económica específica de los contribuyentes.

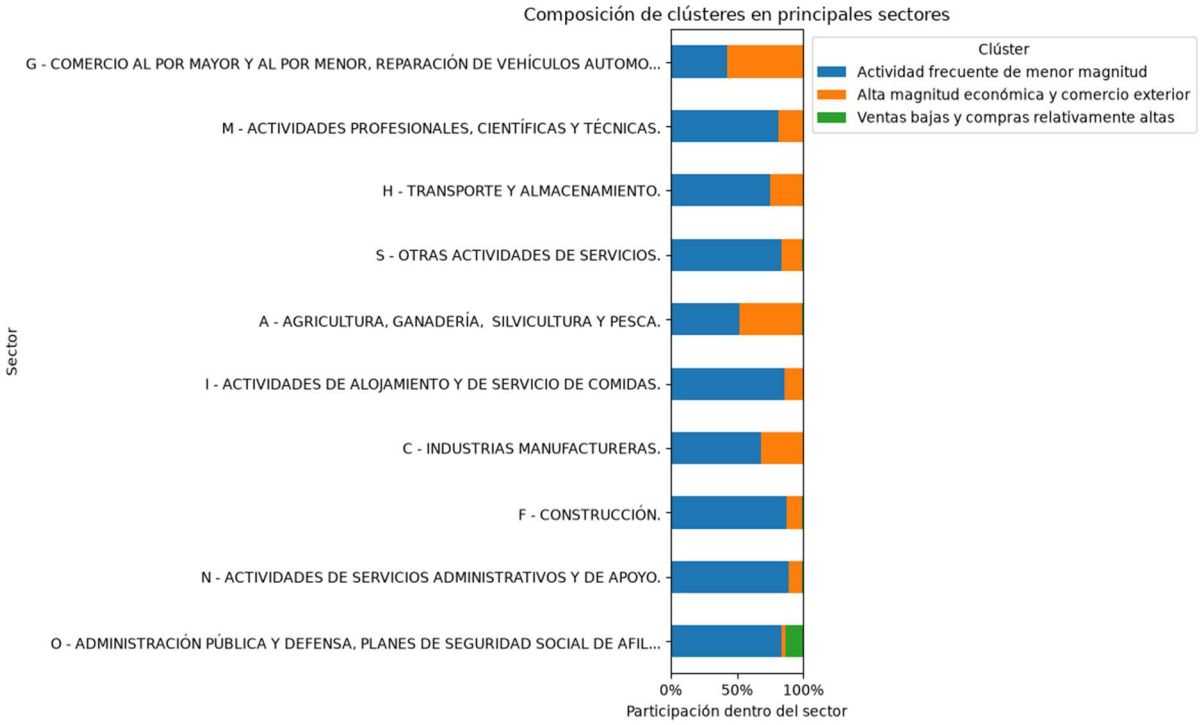
La composición territorial evidencia que la distribución de los perfiles no es homogénea entre provincias y cantones. Entre las provincias con mayor número de registros predomina el clúster Actividad frecuente de menor magnitud, aunque con intensidades distintas. Su participación alcanza 92,88 % en Morona Santiago, 89,34 % en Loja y 87,03 % en Chimborazo. En contraste, el clúster Alta magnitud económica y comercio exterior presenta una mayor

presencia relativa en Pichincha (38,12 %), Guayas (25,23 %), Tungurahua (21,22 %) y El Oro (20,44 %). Al descender al nivel cantonal, este último perfil representa 82,88 % de los registros en Guayaquil, 80,50 % en Cuenca y 77,48 % en Quito. Estos resultados muestran que la agregación provincial puede ocultar diferencias internas y justifican complementar la lectura de los montos territoriales con la composición de los clústeres.

La dimensión sectorial presenta una heterogeneidad similar. En varios de los sectores con mayor número de registros predomina ampliamente el clúster Actividad frecuente de menor magnitud. Este grupo alcanza 88,85 % en Actividades de servicios administrativos y de apoyo, 87,03 % en Construcción y 86,22 % en Actividades de alojamiento y de servicio de comidas. En cambio, Comercio al por mayor y al por menor presenta una estructura distinta: 57,41 % de sus registros corresponde al clúster Alta magnitud económica y comercio exterior. Agricultura, ganadería, silvicultura y pesca muestra una distribución más equilibrada entre este perfil (47,66 %) y el de actividad frecuente (51,62 %). Administración pública y defensa constituye además un caso particular por la presencia del clúster 0, que alcanza 13,71 % de los registros del sector. La Figura 11 sintetiza estos contrastes y muestra que la composición de los perfiles varía según la actividad económica.

Figura 11

Composición porcentual de los clústeres en los principales sectores económicos



Nota. Fuente: Resultados sectoriales del Notebook 04 a partir de datos abiertos del SRI, 2017–2025. Elaboración: propia.

En la dimensión temporal, la persistencia de los tres perfiles entre 2017 y 2025 aporta estabilidad descriptiva, aunque no prueba estabilidad de sujetos individuales. Los registros representan combinaciones agregadas y pueden cambiar de composición cada año. La evolución indica que el clúster de alta magnitud incrementa su participación en registros, mientras que el grupo frecuente disminuye gradualmente; esta tendencia merece seguimiento, pero no permite identificar causas sin variables y diseños adicionales.

Además de las diferencias en la composición de los clústeres, los montos agregados muestran una concentración territorial y sectorial marcada. Pichincha y Guayas reúnen 82,71 % de la recaudación, mientras que comercio lidera el monto sectorial acumulado. Sin embargo, la mayor recaudación por registro se observa en actividades financieras. Estos indicadores

responden a preguntas diferentes y no deben usarse como sinónimos de productividad, rentabilidad, presión tributaria o cumplimiento. Su utilidad consiste en orientar comparaciones y preguntas posteriores.

El dashboard mejora la comunicación porque organiza tareas concretas: resumir, comparar en el tiempo, explorar territorios y examinar perfiles. Esta lógica coincide con Munzner (2014), para quien la representación debe relacionarse con los datos y las tareas del usuario. El dashboard no recalcula el modelo y, por tanto, conserva la separación entre resultados analíticos y presentación interactiva.

En conjunto, el análisis demuestra que una solución técnicamente útil no depende solo del algoritmo. Depende de la calidad de las fuentes, la definición de la unidad de análisis, la selección de variables, la comparación transparente de condiciones computacionales y la prudencia interpretativa. El resultado es adecuado para describir patrones agregados y apoyar la exploración fiscal, no para automatizar decisiones ni atribuir conductas a contribuyentes.

Capítulo 5. Conclusiones, limitaciones y recomendaciones

5.1 Conclusiones

La primera conclusión es que los datos abiertos tributarios del SRI pueden utilizarse para construir un proceso analítico reproducible, siempre que antes se revisen su estructura, calidad y compatibilidad. La apertura de los archivos no elimina la necesidad de diagnóstico; más bien, exige documentar cómo se transforman los datos en una base analítica confiable.

La segunda conclusión es que la preparación de datos fue una etapa decisiva. El tratamiento de valores faltantes, duplicados, ceros, categorías, variables derivadas, escalamiento y selección de variables permitió pasar de fuentes heterogéneas a una base de 391.774 registros apta para modelado no supervisado.

La tercera conclusión es que la comparación de modelos debe combinar métricas internas con interpretación. En este trabajo, la solución K-Means sobre la base reducida, con $k = 3$, fue seleccionada porque ofreció una solución aplicable a toda la base, estable para el propósito del estudio y comprensible para la comunicación de resultados. Las métricas aportaron evidencia, pero no reemplazaron el análisis sustantivo.

La cuarta conclusión es que los clústeres identificados describen patrones agregados diferenciados. El grupo de actividad frecuente concentra la mayor cantidad de registros, mientras que el grupo de alta magnitud económica y comercio exterior concentra la mayor parte de los montos. Esta lectura confirma que frecuencia y magnitud económica son dimensiones distintas del fenómeno analizado. De manera complementaria, el contraste con la clasificación institucional de grandes contribuyentes mostró que el 99,20 % de la recaudación asociada a este segmento durante 2021–2025 se concentró en el clúster de alta magnitud, aportando evidencia externa sobre su coherencia descriptiva, sin implicar equivalencia entre ambas clasificaciones.

La quinta conclusión es que el dashboard en Streamlit convierte los resultados del proceso en una herramienta de exploración coherente y reproducible. Sus cuatro páginas organizan indicadores, evolución temporal, concentración territorial y perfiles económicos, y permiten comparar frecuencia y magnitud sin recalcular el modelo.

La sexta conclusión es que los resultados presentan concentraciones territoriales y sectoriales relevantes. Pichincha y Guayas reúnen el 82,71 % de la recaudación, mientras que comercio, actividades financieras e industrias manufactureras encabezan los montos sectoriales. Estas concentraciones deben interpretarse como patrones agregados del período y no como evidencia sobre contribuyentes individuales.

5.2 Limitaciones

La principal limitación proviene del nivel agregado de los datos. La base no permite analizar contribuyentes individuales ni emitir conclusiones sobre evasión, incumplimiento o riesgo fiscal individual. Los clústeres describen combinaciones agregadas de tiempo, territorio, sector y variables monetarias.

Otra limitación corresponde a la compatibilidad entre fuentes. Aunque se construyó una integración preliminar sobre dimensiones comunes, algunas variables pertenecen únicamente a una fuente o presentan niveles de detalle distintos. Integrarlas sin cuidado habría generado duplicaciones o interpretaciones artificiales.

También existe una limitación computacional. K-medoids y agrupamiento jerárquico fueron evaluados sobre muestras reproducibles debido a su mayor costo en bases grandes. Por ello, sus resultados fueron útiles para comparación, pero no reemplazaron la necesidad de seleccionar un modelo aplicable a la totalidad de registros.

Finalmente, el dashboard se plantea como prototipo académico. Una versión institucional requeriría pruebas adicionales de rendimiento, seguridad, actualización de datos, control de versiones, experiencia de usuario y gobierno de datos.

5.3 Recomendaciones

Se recomienda conservar la trazabilidad completa de archivos originales, notebooks, scripts, parámetros, bases procesadas y tablas finales. Esta práctica facilita la revisión, la reproducción y la corrección de errores.

Se recomienda conservar una guía breve de uso, capturas actualizadas de las cuatro páginas y un procedimiento documentado para regenerar las tablas mediante los Notebooks 01 a 04. También conviene validar los filtros después de cada actualización de datos y limpiar la caché de Streamlit para asegurar que la aplicación muestre los archivos recién generados.

Para trabajos futuros, se podrían explorar modelos adicionales, estrategias de reducción de dimensionalidad o comparaciones por subperíodos, sectores específicos y territorios de interés. Estas extensiones deberían mantener el mismo cuidado sobre granularidad e interpretación agregada.

Cualquier uso institucional de este tipo de análisis debería incorporar revisión experta, criterios éticos, gobierno de datos y supervisión humana. La analítica puede apoyar la comprensión y priorización, pero no debe transformarse en un mecanismo automático de decisión sin controles adecuados.

Referencias

Aslett, J., González, G., Hamilton, S., & Pecho, M. (2024). *Tax administration: Essential analytics for compliance risk management* (IMF Technical Notes and Manuals 2024/001). International Monetary Fund. https://doi.org/10.5089/9798400260063.005

Aslett, J., González, I., Hadwick, D., Hamilton, S., Hardy, M. A., & Pérez, A. (2024). *Understanding artificial intelligence in tax and customs administration* (IMF Technical Notes and Manuals 2024/006). International Monetary Fund. https://doi.org/10.5089/9798400290435.005

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), Article 16. https://doi.org/10.1145/1541880.1541883

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), 1-27. https://doi.org/10.1080/03610927408827101

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224-227. https://doi.org/10.1109/TPAMI.1979.4766909

de Roux, D., Pérez, B., Moreno, A., Del Pilar Villamil, M., & Figueroa, C. (2018). Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 215-222). Association for Computing Machinery. https://doi.org/10.1145/3219819.3219878

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International*

Conference on Knowledge Discovery and Data Mining (pp. 226-231). AAAI Press. https://aaai.org/papers/kdd96-037-a-density-based-algorithm-for-discovering-clusters-in-large-spatial-databases-with-noise

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54. https://doi.org/10.1609/aimag.v17i3.1230

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management, 29*(4), 258-268. https://doi.org/10.1080/10580530.2012.716740

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley. https://doi.org/10.1002/9780470316801

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281-297). University of California Press. https://digicoll.lib.berkeley.edu/record/113015

Munzner, T. (2014). *Visualization analysis and design*. CRC Press. https://doi.org/10.1201/B17511

OECD. (2016). *Advanced analytics for better tax administration: Putting data to work*. OECD Publishing. https://doi.org/10.1787/9789264256453-en

OECD. (2024). *Tax administration 2024: Comparative information on OECD and other advanced and emerging economies*. OECD Publishing. https://doi.org/10.1787/2d5fba9c-en

Ordoñez, J., & Hallo, M. (2023). Detección de comportamientos atípicos de contribuyentes con riesgo de no pago en una administración tributaria: Un marco de trabajo de minería de datos. *Revista Politécnica, 52*(1), 35-44. https://doi.org/10.33333/rp.vol52n1.04

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Savić, M., Atanasijević, J., Jakovetić, D., & Krejić, N. (2022). Tax evasion risk management using a hybrid unsupervised outlier detection method. *Expert Systems with Applications, 193*, 116409. https://doi.org/10.1016/j.eswa.2021.116409

Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems, 101*, 101804. https://doi.org/10.1016/j.is.2021.101804

Servicio de Rentas Internas. (s. f.). *Datasets.* Recuperado el 19 de julio de 2026, de https://www.sri.gob.ec/datasets

Servicio de Rentas Internas. (2021). *Resolución Nro. NAC-DGERCGC21-00000022: Expedir los criterios para considerar a sujetos pasivos como grandes contribuyentes para fines tributarios.* https://www.sri.gob.ec/o/sri-portlet-biblioteca-alfresco-internet/descargar?id=f7876a79-cf7b-4612-9e7a-6939485377ac&nombre=NAC-DGERCGC21-00000022.pdf

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33. https://doi.org/10.1080/07421222.1996.11518099

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29-39). https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Xu, R., & Wunsch, D. C., II. (2009). *Clustering*. Wiley-IEEE Press. https://doi.org/10.1002/9780470382776

Zheng, Q., Xu, Y., Liu, H., Shi, B., Wang, J., & Dong, B. (2024). A survey of tax risk detection using data mining techniques. *Engineering, 34*, 43-59. https://doi.org/10.1016/j.eng.2023.07.014

Anexos

Anexo A. Relación entre notebooks y entregables

Tabla 12

Trazabilidad de notebooks y entregables principales

Notebook	Propósito	Entregables principales
Notebook 01	Extracción, diagnóstico e integración preliminar	Inventario de archivos, bases consolidadas, base integrada preliminar y reportes de integración
Notebook 02	Preparación de variables y dataset analítico final	Base de modelado tratada, matrices escaladas, variables finales y bases PCA
Notebook 03	Modelado no supervisado y evaluación	Etiquetas de modelos, métricas comparativas, selección de K-Means k = 3 y tablas para dashboard
Notebook 04	Resultados y preparación del dashboard en Streamlit	Indicadores, gráficos, tablas sectoriales enriquecidas, estructura del dashboard e inventarios finales
Notebook 05	Implementación del dashboard en Streamlit	Estructura de aplicación, módulos de carga de datos, app principal y páginas temáticas

Nota. Fuente: Notebooks 01 a 05 y estructura de entregables del proyecto. Elaboración: propia.

Anexo B. Archivos principales generados

base_integrada_preliminar.parquet

base_modelado_tratada.parquet

base_clustering_completa_escalada.parquet

base_clustering_reducida_escalada.parquet

dashboard_base_clusters.csv

dashboard_temporal_clusters.csv

dashboard_provincia_clusters.csv

dashboard_canton_clusters.csv

dashboard_sector_clusters_con_nombres.csv

dashboard_resumen_clusters_presentacion.csv

dashboard_resumen_sectores_con_nombres.csv

dashboard_estructura_prevista.csv

dashboard_estructura_datos_streamlit.csv

dashboard_criterios_visualizacion.csv

Anexo C. Nota de reproducibilidad y alcance de los resultados

Los resultados del documento, las tablas finales y el dashboard se derivan del mismo flujo de cinco notebooks. Para reproducir el producto se deben conservar los archivos originales, ejecutar los notebooks en orden, verificar la creación de los CSV requeridos y reiniciar la caché de Streamlit. Las semillas aleatorias y los reportes de validación reducen variaciones no controladas, aunque cambios de versión en las librerías pueden producir pequeñas diferencias numéricas.

Las etiquetas de clúster son descriptivas y dependen de la selección de variables, el escalamiento y la configuración del modelo. No representan categorías oficiales ni diagnósticos sobre contribuyentes. Cualquier actualización de las fuentes debe repetir las validaciones de estructura, calidad, métricas e interpretación antes de comparar resultados con esta versión.