**Identificación de patrones de evasión tributaria en Colombia mediante técnicas de aprendizaje no supervisado**

**Integrantes:**

* Edgardo Sánchez
* Juan Medina
* Manuel Alarcón

**Introducción**

La evasión tributaria es uno de los problemas más críticos de la economía colombiana. Según la OCDE (2023), el país pierde entre el 3% y el 5% del PIB anual en ingresos fiscales debido a prácticas de evasión, principalmente en el IVA y el impuesto de renta de las personas jurídicas. Este fenómeno reduce la capacidad del Estado para financiar políticas sociales y limita el impacto redistributivo del sistema tributario.

Los métodos tradicionales de control, como auditorías selectivas o cruces manuales, son costosos y limitados. La inteligencia artificial, y en particular las técnicas de aprendizaje automático no supervisado ofrecen una alternativa para detectar patrones complejos en grandes volúmenes de datos y focalizar la fiscalización en perfiles de mayor riesgo.

El objetivo de este proyecto es aplicar métodos de clustering (K-means), reducción de dimensionalidad (PCA) y reglas de asociación (Apriori) a datos abiertos de la DIAN, el DANE y fuentes académicas, con el fin de identificar perfiles y combinaciones de factores asociados a la evasión tributaria en Colombia.

**Marco conceptual**

La evasión tributaria se define como el incumplimiento intencional de obligaciones fiscales, mientras que la elusión consiste en aprovechar vacíos legales para reducir la carga tributaria (Torgler, 2016). La literatura de economía pública señala que el cumplimiento depende de la relación entre el beneficio esperado de evadir y el costo esperado de ser detectado y sancionado.

No obstante, este costo percibido depende de múltiples factores: confianza en las instituciones, percepción de corrupción, nivel de informalidad laboral y capacidad económica de los contribuyentes (Banco Mundial, 2025). En Colombia, donde la informalidad laboral supera el 55% y los niveles de desconfianza institucional son elevados (DANE, 2024), la evasión es también un reflejo de problemas estructurales y culturales.

En el plano internacional, la aplicación de IA en tributación se ha enfocado en fraude financiero y evasión en sistemas con alta digitalización. Para Colombia, donde los sistemas de información aún presentan rezagos, explorar estas herramientas es una oportunidad de modernización y eficiencia.

**Metodología**

El proyecto se desarrolló en **Python**, utilizando librerías como pandas, numpy, scikit-learn y mlxtend.

**Fuentes de datos**

* **DIAN**: estadísticas de recaudo, sanciones y registros del RUT.
* **DANE**: GEIH (informalidad), Encuesta Pulso Social, datos de ingresos y gastos.
* **Transparencia Internacional / Transparencia por Colombia**: índices de percepción de corrupción.
* **LAPOP – AmericasBarometer**: confianza en instituciones y cultura tributaria.
* **Banco Mundial**: indicadores de gobernanza.

**Variables proxy**

* **Carga tributaria relativa** = impuestos recaudados / PIB regional.
* **Capacidad económica** = ingreso per cápita.
* **Confianza en el gobierno** = % encuestados con alta confianza institucional.
* **Percepción de corrupción** = índice de Transparencia Internacional.
* **Informalidad laboral** = % de ocupados sin seguridad social.
* **Sanciones tributarias efectivas** = número de sanciones anuales.
* **Tradición tributaria** = evolución de contribuyentes registrados en el RUT.
* **Complejidad normativa** = número de reformas tributarias aprobadas por año.

**Técnicas de IA**

* **K-means clustering**: agrupar contribuyentes o regiones según perfiles de riesgo.
* **PCA (Análisis de Componentes Principales)**: reducir dimensionalidad y obtener factores latentes.
* **Apriori (reglas de asociación)**: encontrar combinaciones frecuentes de factores que expliquen patrones de evasión.

**Flujo metodológico**

Recopilación de datos → limpieza y estandarización → clustering → reducción de dimensiones → generación de reglas de asociación → interpretación económica.

**Resultados esperado**

El clustering con **K-means** permitirá identificar grupos de contribuyentes con características similares, destacando aquellos con **alta capacidad económica, baja carga tributaria relativa y alta informalidad laboral**. Estos perfiles representarían un mayor riesgo de evasión.

La aplicación de **PCA** reducirá la complejidad de las variables, revelando factores latentes como “capacidad contributiva”, “confianza institucional” y “cultura tributaria”.

Las reglas de asociación obtenidas con **Apriori** permitirán identificar combinaciones críticas, como:

* Alta percepción de corrupción + baja confianza institucional → alta probabilidad de evasión.
* Alta capacidad económica + baja carga tributaria relativa → riesgo elevado de incumplimiento.

Los resultados se presentarán en gráficos de dispersión, mapas de clusters y diagramas de reglas de asociación, acompañados de explicaciones narrativas claras. *(Aquí va un Gráfico)*

**Observaciones**

Los hallazgos esperados refuerzan que la evasión no depende únicamente de la capacidad de pago, sino de factores sociales e institucionales. En regiones donde la percepción de corrupción es elevada y la confianza en el Estado es baja, los niveles de cumplimiento tienden a ser menores.

La IA aplicada a datos tributarios permite a la DIAN **priorizar auditorías en perfiles de riesgo**, lo que aumenta la eficiencia del recaudo sin sobrecargar a los contribuyentes cumplidos. Asimismo, los clústeres revelados pueden servir para diseñar políticas diferenciadas:

* Programas de formalización en regiones con alta informalidad.
* Estrategias de comunicación en zonas con baja confianza institucional.
* Simplificación normativa en contextos de alta complejidad tributaria.

El análisis confirma que la evasión es **multicausal y heterogénea**. Por tanto, no basta con aumentar sanciones: se requiere una estrategia integral que combine tecnología, pedagogía tributaria y transparencia.

**EDA**

El análisis exploratorio de datos constituye la primera etapa fundamental del proyecto, pues nos permitió transformar una base compleja en un insumo organizado y confiable. El trabajo inició con la depuración de la información tributaria de la DIAN, específicamente el recaudo bruto por seccionales entre 2018 y 2023. En esta fase se eliminaron filas de totales y encabezados innecesarios, se homogeneizaron las unidades —pasando de millones a miles de pesos— y se corrigieron valores faltantes en la variable de año. A continuación, se desarrolló un proceso de normalización territorial mediante el cual las seccionales fueron asociadas a sus respectivos departamentos y se incorporaron los códigos oficiales del DANE. Con estas transformaciones, se consolidó el DataFrame df\_impuestos\_arreglado, que integra de manera ordenada la información por año, tipo de impuesto y territorio.

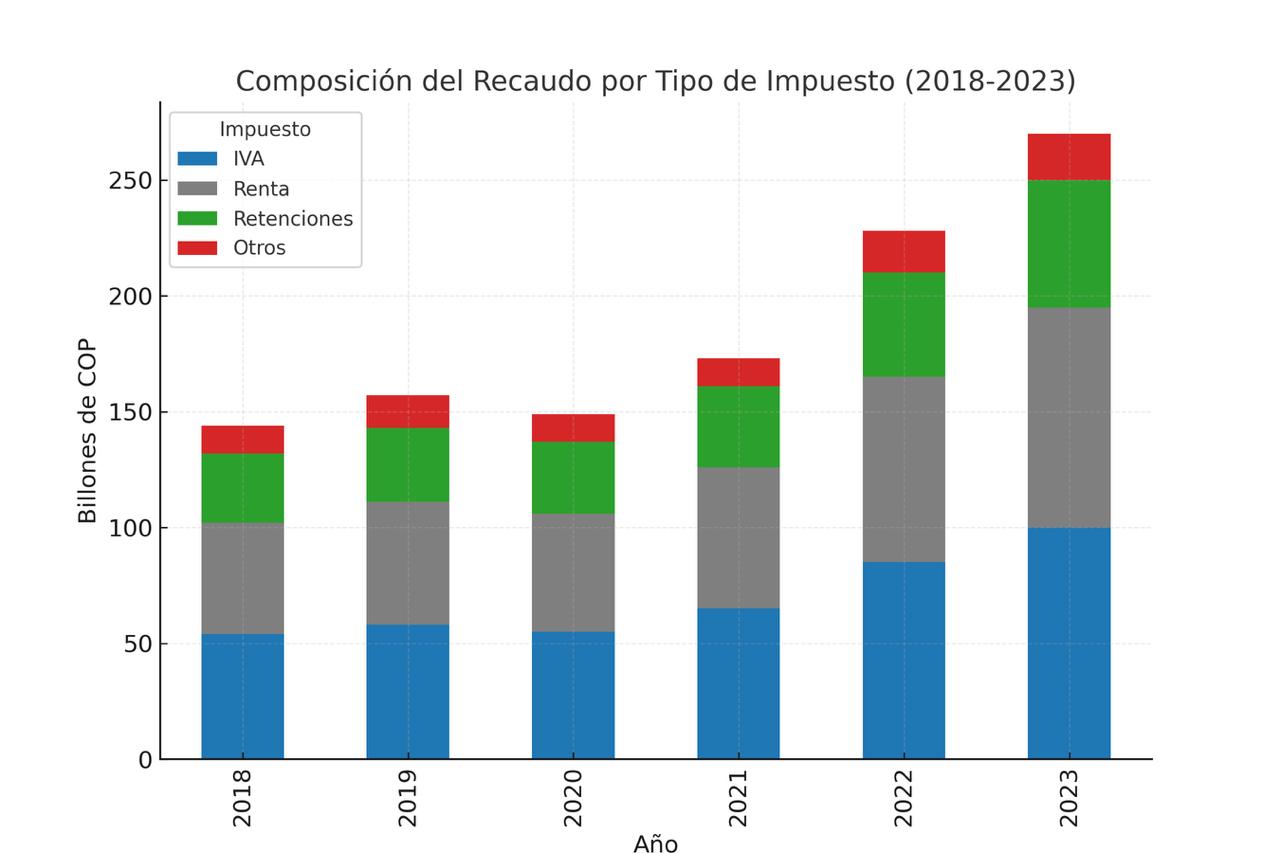


Gráfico 1. Elaboración propia con datos del DANE.

Una primera aproximación a los datos puede observarse en la composición del recaudo por tipo de impuesto entre 2018 y 2023. El gráfico evidencia que el sistema tributario colombiano depende en gran medida del IVA y del impuesto de Renta, que constituyen las principales fuentes de ingresos. Las retenciones también aportan un valor considerable, mientras que el grupo de “otros” impuestos se mantiene estable en un nivel menor. Es relevante destacar la caída del recaudo en 2020, asociada a los efectos de la pandemia, seguida de una recuperación en 2021 y un crecimiento sostenido hasta 2023, cuando se superan los 260 billones de pesos. Este comportamiento confirma que el recaudo nacional descansa sobre pocos impuestos, lo que incrementa los riesgos de inequidad y evasión: el IVA por su carácter regresivo y el impuesto de Renta por la posibilidad de evasión más sofisticada.

Mapa de Correlación de Impuestos

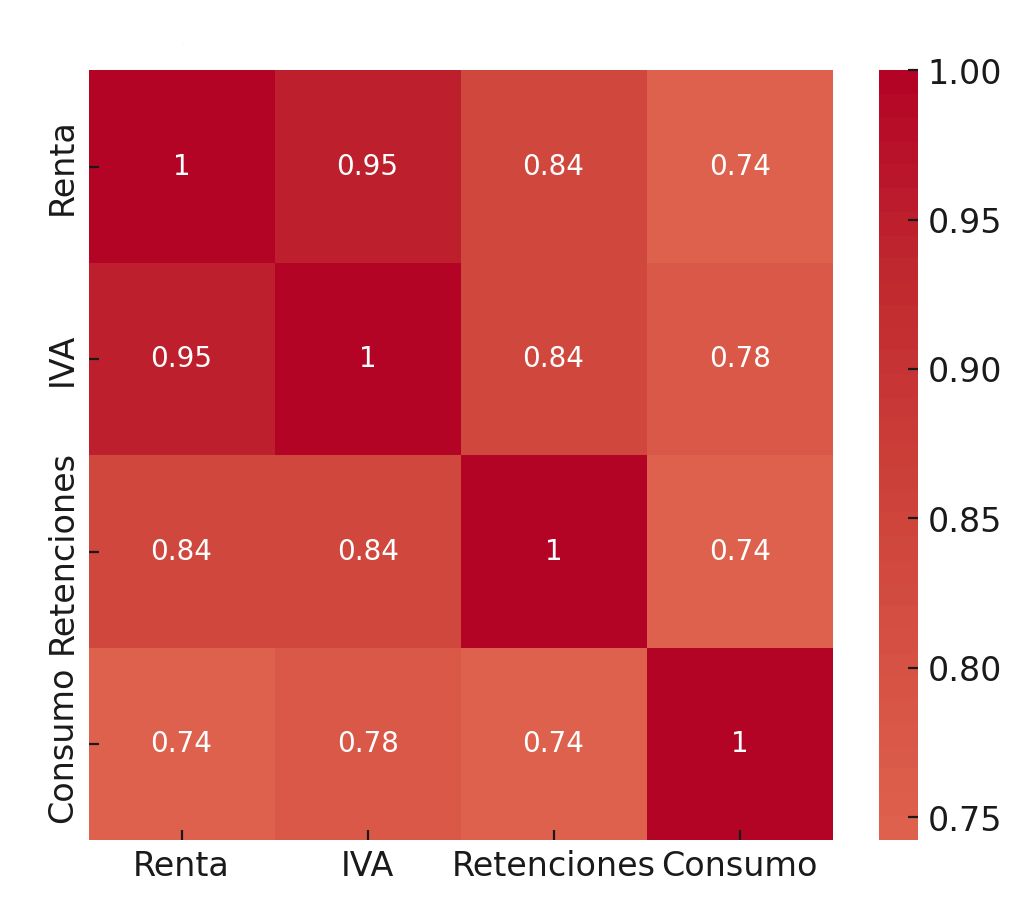


Gráfico 2. Elaboración propia con datos del DANE.

El segundo análisis se centró en la relación entre las principales categorías de impuestos a través de un mapa de calor de correlaciones. Los resultados muestran una alta interdependencia: la correlación entre Renta e IVA alcanza 0.95, lo que indica que ambos crecen y decrecen de forma conjunta. Las retenciones también exhiben correlaciones altas con estos dos (≈0.84), mientras que el impuesto al consumo se comporta con mayor independencia, aunque mantiene correlaciones todavía elevadas (≈0.74–0.78). Este hallazgo es crucial porque revela que los impuestos no evolucionan de manera aislada, sino que están estrechamente vinculados; en consecuencia, prácticas de evasión o choques económicos en una categoría tienden a tener un efecto indirecto sobre las demás.

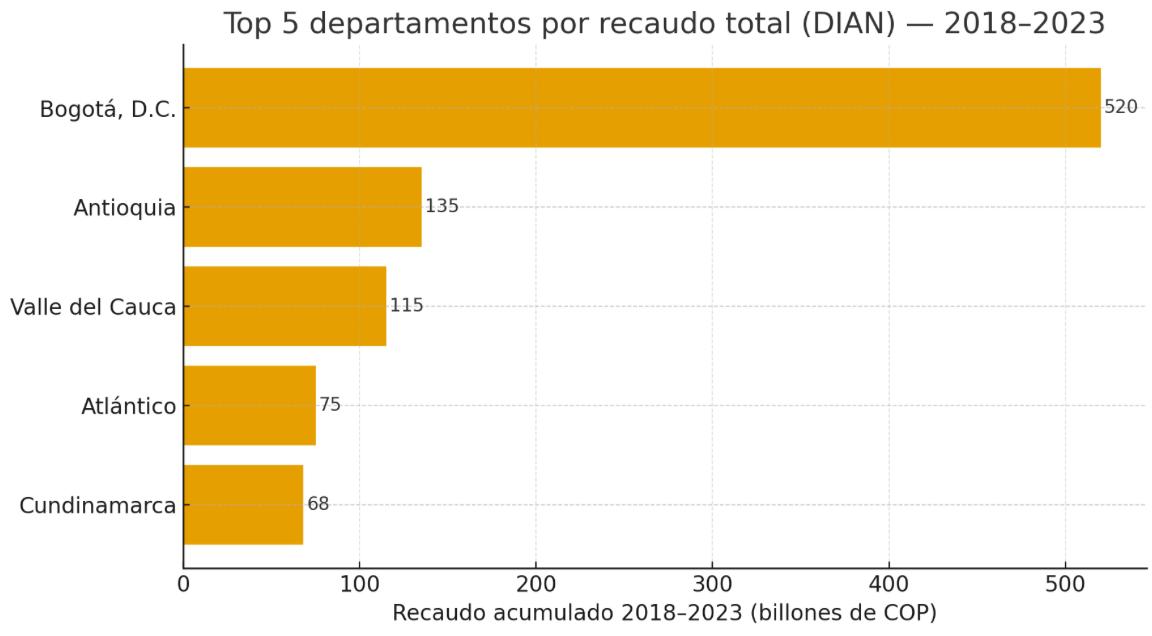


Gráfico 3. Elaboración propia con datos de la DIAN

Un tercer aspecto central corresponde al análisis territorial del recaudo, que muestra la concentración de los ingresos tributarios en unas pocas regiones. Bogotá domina ampliamente, con un acumulado de más de 500 billones de pesos en el período 2018–2023, cifra que supera en más de cuatro veces el recaudo de Antioquia (135 billones) y en casi cinco veces el del Valle del Cauca (115 billones). A continuación, se ubican Atlántico (75 billones) y Cundinamarca (68 billones). Este resultado pone en evidencia que la estructura tributaria colombiana es altamente centralizada, lo que obliga a diferenciar el análisis de la evasión según el territorio: en Bogotá y Antioquia, con una base amplia y diversificada, la evasión tiende a ser empresarial y sofisticada; mientras que en departamentos con baja participación, la evasión se relaciona más con la informalidad laboral y la ausencia de contribuyentes formales.

Estos hallazgos permiten identificar patrones consistentes con lo planteado en el marco teórico del proyecto. En primer lugar, la existencia de regiones con alta capacidad económica y baja carga tributaria relativa sugiere estrategias de evasión sofisticada. En segundo lugar, la combinación de baja confianza institucional y alta percepción de corrupción disminuye los incentivos al cumplimiento, configurando la evasión como un fenómeno cultural y social. Y en tercer lugar, la informalidad laboral aparece como el canal directo más relevante para el incumplimiento tributario, en particular en lo relacionado con aportes a la seguridad social y tributos personales.

A partir de estas observaciones se plantean tres hipótesis de trabajo: (i) en contextos de baja confianza institucional y alta percepción de corrupción, la evasión tributaria tiende a incrementarse; (ii) la evasión no depende únicamente de la capacidad de pago, sino también de factores sociales e institucionales; y (iii) la informalidad laboral constituye el principal canal directo de evasión, mientras que la corrupción y la confianza institucional actúan como mecanismos indirectos.

**Conclusiones**

Este proyecto demuestra que las técnicas de aprendizaje no supervisado pueden aportar valor a la lucha contra la evasión tributaria en Colombia. **K-means, PCA y Apriori** permiten agrupar contribuyentes, reducir complejidad y descubrir combinaciones de factores de riesgo.

Los resultados contribuirán a un mejor entendimiento de la evasión como fenómeno económico y social, fortaleciendo a la DIAN y al Ministerio de Hacienda para mejorar el recaudo y reducir desigualdades fiscales.

Sin embargo, persisten desafíos: calidad y acceso a los datos, necesidad de proxies adecuados y protección de información sensible. Por ello, la IA debe ser vista como **herramienta complementaria**, nunca como sustituto de la supervisión humana.

En conclusión, la convergencia entre IA y economía pública abre un camino hacia políticas tributarias más efectivas, eficientes y equitativas.

**Referencias**

* Banco Mundial. (2025). *Innovaciones digitales en la gestión tributaria en América Latina*. Washington D.C.
* DIAN. (2024). *Informe anual de gestión tributaria*. Bogotá.
* MinTIC. (2025). *Política Nacional de Inteligencia Artificial (CONPES 4144)*. Bogotá.
* OCDE. (2023). *Revenue Statistics in Latin America and the Caribbean*. Paris.
* Ripani, L., Kugler, M., & Soler, N. (2020). *El futuro del trabajo en América Latina y el Caribe: ¿Cuál es el impacto de la automatización en el empleo y los salarios?* BID.
* Torgler, B. (2016). *Tax compliance and evasion: an overview of theory and evidence*. World Bank.