Algoritmos para la fiscalización inteligente en el Valle del Cauca

Algorithms for intelligent inspection in the Valle del Cauca

Algoritmos para inspeção inteligente no Valle del Cauca

Sonia Castro Y.*
Liliana Plaza Ñ.**
Luis Carlos Torres S.***

* Contadora pública; magíster en Económica, Universidad Javeriana. Becaria MinTic en la especialización en Analítica de Negocios, ICESI. scastro@valledelcauca.gov.co / https://orcid.org/0000-0003-1239-5268

** Economista con profundización en Econometría; especialista en Mercados y Políticas del Suelo de América Latina, Universidad Nacional de Colombia; magíster en Economía de la Universidad Manizales. lmplaza@valledelcauca.gov.co / https://orcid.org/0000-0002-4233-4207

*** Matemático y magíster en Ingeniería de Sistemas, Universidad Nacional de Colombia; maestría en Ciencias de la Educación, Universidad de Sherbrooke (Canadá); doctorado Pensamiento Complejo, Multiversidad Mundo Real Edgar Morin, México. Autor de: *Problemas para la inteligencia artificial y natural* y *Creatividad en el aula*. Docente-investigador en la Universidad El Bosque. lctorress@gmail.com / http://orcid.org/0000-0001-6756-4984.
DOI: https://doi.org/10.18601/16926722.n20.07

Resumen

La escasez de herramientas para detectar contribuyentes que no cumplen con sus obligaciones tributarias, y la imposibilidad de generar planes de fiscalización presenciales con la actual situación de aislamiento por la pandemia de covid-19, debilitan la generación de valor a las entidades recaudadores, que tienen como fin recaudar recursos para la inversión social como recreación y salud, o para escuelas, puentes y hospitales. Con esta problemática, la Secretaría de las Tecnologías de la Información y Comunicaciones del Valle del Cauca diseñó estrategias basadas en evidencia, con datos de la Unidad de Rentas y de las cámaras de comercio y la DIAN del Valle del Cauca. Con esta información se procedió a crear algoritmos para la predicción de contribuyentes omisos de los impuestos departamentales, al igual que programas con georreferenciación. Se analizan los estados financieros de las 4.525 empresas que reportan a la siete cámaras de comercio del departamento del Valle del Cauca y la base de datos de la DIAN con 686.215 datos de sus reportes de la declaración anual de impuesto mínimo alternativo simple (IMAS) para trabajadores por cuenta propia; IMAS para empleados; declaración de renta y complementarios personas naturales y asimiladas de residentes y sucesiones ilíquidas de causantes residentes; declaración de renta y complementarios o de ingresos y patrimonio para personas jurídicas y asimiladas, y personas naturales y asimiladas no residentes y sucesiones ilíquidas de causantes no residentes. El procedimiento para el análisis de los dos entes, tanto de la DIAN como de las Cámaras, se realizó por separado en el año 2018. La evidencia y la metodología propuestas presentan una gran pertinencia para las políticas públicas basadas en algoritmos para la focalización de la fiscalización.

Palabras clave: focalización de políticas públicas; tributaria; algoritmos; cruce de información.

Abstract

The scarcity of tools to detect taxpayers who do not meet their tax obligations and the inability to generate face-to-face control plans with the current situation of isolation by the Covid pandemic, weaken the generation of value to collecting entities, which are intended to raise resources for social investment, such as recreation, health, or in schools, bridges, hospitals. With this problem, the secretary oInformation and Communication Technologies (ICTs) of the Valle del Cauca, designed evidence-based strategies, with data, both internal, from the Revenue Unit and exogenous as well as those reported by the Chambers of Commerce and the DIAN of Valle del Cauca. This information resulted in the creating of algorithms for the prediction of omissive contributors to departmental taxes as well as programs with georeferencing. It analyzes the financial statements of the 4,525 companies reporting to the seven trading chambers of the Valle del Cauca

department and the DIAN database with 686,215 data, from their reports of the Annual Simple Alternative Minimum Tax Return (IMAS) for self-employed, Annual Alternative Minimum Tax Return for Employees (IMAS), Declaration of Income and Complementary Natural and Assimilated Persons of Residents and Líquid Successions of Resident Causes, Declaration of Income and Supplemental or Income and Heritage for Legal and Assimilated Persons and Natural and Assimilated Non-Resident Persons and Líquid Successions of Non-Resident Causes. The procedure for the analysis of both the DIAN and the Chambers was carried out separately in 2018. The evidence and proposed methodology are of great relevance to algorithm-based public policies for the targeting of auditing.

Key words: Public policy targeting; taxation; algorithms; information crossover.

Resumo

A escassez de ferramentas de detecção de contribuintes que descumpram suas obrigações tributárias e a impossibilidade de gerar planos de fiscalização face a face com a atual situação de isolamento devido à pandemia de Covid fragilizam a geração de valor para as entidades arrecadadoras, cujo objetivo é arrecadar recursos para investimento social, como recreação, saúde, ou em escolas, pontes, hospitais. Com este problema, a Secretaria de Tecnologias da Informação e Comunicações do Valle del Cauca traçou estratégias baseadas em evidências, com dados, tanto internos, da Unidade da Receita, quanto exógenos, que foram reportados pelas Câmaras de Comércio e DIAN de Valle del Valle Cauca. Com essas informações, passamos a criar algoritmos para a previsão de contribuintes omitidos dos tributos departamentais e também programas com georreferenciamento. São analisados os demonstrativos financeiros das 4.525 empresas que reportam às sete câmaras de comércio do departamento de Valle del Cauca e a base de dados DIAN com 686.215 dados de seus relatórios do Imposto Anual Simples Alternativo Mínimo (IMAS). para Trabalhadores Autônomos, Declaração Anual de Imposto Mínimo Alternativo (IMAS) para Funcionários, Declaração de Renda e Pessoas Naturais e Assimiladas Complementares de Residentes e Sucessões Ilíquidas de Causadores Residentes, Declaração de Renda e Complementares ou Rendimentos e Ativos para Pessoas Jurídicas Pessoas Naturais e Assimiladas Assimiladas e Não Residentes e Sucessões Ilíquidas de Causadores Não Residentes. O procedimento de análise das duas entidades, DIAN e Câmaras, foi realizado separadamente em 2018. As evidências e a metodologia proposta são de grande relevância para as políticas públicas baseadas em algoritmos de direcionamento à auditoria.

Palavras-chave: Direcionamento de políticas públicas; impostos; algoritmos; cruzamento de informações.

Introducción

Desde 2019, la Gobernación del Valle está creando prototipos que ayuden a generar valor y conocimiento con la analítica de los datos; para esto, en coordinación con la Unidad Administrativa Especial de Rentas y Gestión Tributaria, realizó una propuesta para incrementar los ingresos y tener una gestión eficiente, transversal y, sobre todo, más equitativa, en la cual los obligados a pagar impuestos cumplirían su responsabilidad de contribuir[1] al Estado para los fines comunes del departamento del Valle del Cauca. Así, la investigación tiene como fin crear una herramienta para la detección de contribuyentes omisos de los impuestos departamentales. Su gestión de cobro, su liquidación y los cálculos sobre sanciones, intereses o impuestos están a cargo exclusivo de la Unidad de Rentas del Departamento[2].

Con esta iniciativa, con la ayuda del Ministerio de las Tecnologías de la Información y las Comunicaciones (Mintic) que, a través de una convocatoria, nos facilitó los datos, y con la empresa proveedora de servicios tecnológicos en la Gobernación del Valle se han realizado los análisis a las bases de datos de la DIAN.

El piloto de la analítica de los datos fue realizado en dos grupos y con diferente estructura de *software* para lograr el objetivo planteado. Los resultados fueron socializados a la Unidad de Rentas.

El objetivo del presente trabajo es realizar un programa de inteligencia fiscal[3] con las bases de datos externas en el caso de impuestos departamentales, a saber: contribución para el deporte, la recreación y el aprovechamiento del tiempo libre; impuesto de loterías foráneas; impuesto al degüello de ganado mayor, para la detección de contribuyentes omisos o evasores, con sus respectivos algoritmos y balance Score Card.

Esta investigación se realiza con fundamento en la ciencia de datos, con el principio de relación de causalidad en la actividad económica del contribuyente y la determinación del impuesto[4].

---

1 Ordenanza 474 de la Gobernación del Valle del Cauca, artículo 96, pp. 40-46.

2 Pronunciamiento del Consejo de Estado, Sentencia 2005-01421 de 30 de octubre de 2013.

3 La Inteligencia Fiscal Tributaria tiene como objetivo obtener pruebas que sirvan para validar las inferencias de una fiscalización, es decir, buscar datos que sean el insumo para fiscalizaciones tributarias profundas que permitan iniciar un proceso penal o administrativo y, en lo posible, recuperar la deuda por el no pago de impuestos.

4 El artículo 107 del Estatuto Tributario señala los requerimientos para que las expensas necesarias sean deducibles: "Son deducibles las expensas realizadas durante el año o período gravable en el desarrollo de cualquier actividad productora de renta, siempre que tengan relación de causalidad con las actividades productoras de renta y que sean necesarias y proporcionadas de acuerdo con cada actividad". Se cita el concepto dado por la Dirección de Impuestos y Aduanas Nacionales DIAN, Oficio 1093, con fecha 2018-01-17, para la deducibilidad de gastos o erogaciones en asociaciones sindicales. En la disposición legal citada se consagran los presupuestos esenciales para que los gastos sean deducibles, como son: a) la relación de causalidad b) la necesidad y c) la proporcionalidad. La *relación de causalidad* significa que los gastos, erogaciones o simplemente la salida de recursos deben guardar una relación causal con la actividad u ocupación que le genera la renta al contribuyente. Esa relación, vínculo o correspondencia debe establecerse entre la expensa (costo o gasto) y la actividad que desarrolla el objeto social (principal o secundario) pero que en todo caso le produce renta, de manera que sin aquella no es posible obtener esta, que en términos de otras áreas del derecho se conoce como nexo causal o relación causa-efecto.

Para la determinación de los algoritmos se solicitó el concepto tributario en la Unidad de Rentas (tabla1).

Tabla 1. Relación entre impuestos y actividades económicas

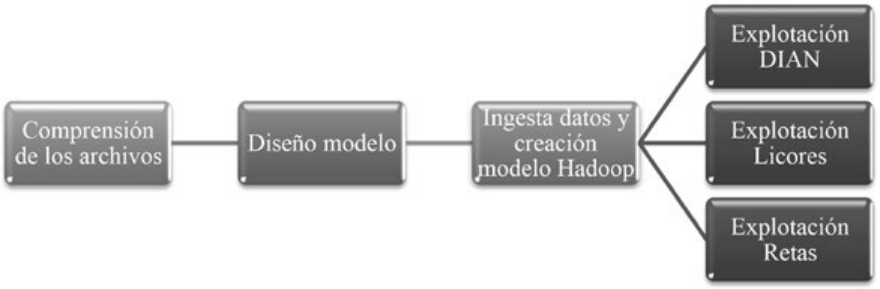
Impuesto	Actividades económicas
Contribución para el deporte, la recreación y el aprovechamiento del tiempo libre	5512, 5530, 5590
Impuesto de loterías foráneas	9200
Impuesto al degüello de ganado mayor	1011

Fuente: Unidad Administrativa Especial de Rentas y Gestión Tributaria.

Una de las principales características de este tipo de trabajos es el manejo de datos sensibles, por lo cual se realizan, tanto para la Dirección de Impuestos y Adunas Nacionales (DIAN) como para las Cámaras de Comercio, protocolos que garanticen el manejo de los mismos. Al respecto, la Secretaría de las Tecnologías de la Información y las Comunicaciones realizó un algoritmo de encriptación de la información sensible de los contribuyentes, esto lleva a anonimizar los números de identificación tributaria (NIT) de los 686.215 datos reportados por la DIAN, y las 4.525 empresas reportadas en la Cámara de Comercio del Valle del Cauca, y se eliminaron los demás datos como teléfono, dirección de los contribuyentes y representante legal.

Se realizó un análisis de cada una de las variables de los formularios de la DIAN, posteriormente se creó un modelo relacional, se incorporaron los programas Impalaprograma[5], Hive[6] y R-studio[7], y se hizo un análisis descriptivo de las variables de la Data Set y su respectiva relación entre los impuestos departamentales. Sin embargo, la etapa de exploración, identificación de métricas de las variables y su minería cubre casi el 70 % del programa.

Figura 1.



Fuente: elaboración propia.

5 Programa funcional para ingesta de datos.

6 Programa funcional para bases de datos.

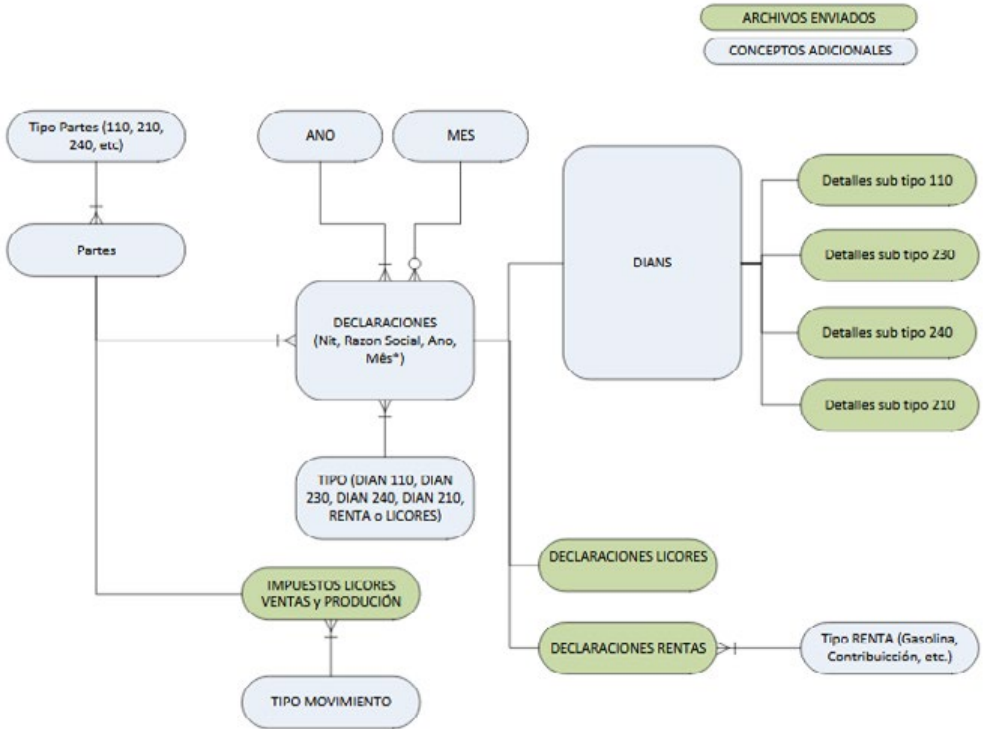
7 Programa para el análisis estadístico.

I. Datos

Las variables integradas al análisis fueron los formatos anual de impuesto mínimo alternativo simple (IMAS) para trabajadores por cuenta propia; Declaración anual de impuesto mínimo alternativo (IMAS) para empleados; Declaración de renta y complementarios personas naturales y asimiladas de residentes y sucesiones ilíquidas de causantes residentes; Declaración de renta y complementario o de ingresos y patrimonio para personas jurídicas y asimiladas y personas naturales y asimiladas no residentes, y sucesiones ilíquidas de causantes no residentes para el año 2018.

Para el análisis se procede a estructurar un Data Set, en el cual se unifican las bases de datos identificando las llaves y las variables en común, y si es necesaria o no su transformación; posteriormente, se realiza un modelo relacional (figura 2).

FIGURA 2. MODELO RELACIONAL DE LA BASE DE DATOS DE LA DIAN



Fuente: Unidad Administrativa Especial de Rentas y Gestión Tributaria.

II. Análisis descriptivo de las variables

Con el modelo relacional y las variables depuradas se procede a realizar los análisis con el programa R para cada una de las variables que se tenían y se crean algoritmos relacionales

y los Plot (uso de librerías de visualización en R) y los algoritmos llamados Random Forest (algoritmos de árboles de decisión); posteriormente, se realizan los algoritmos de asociación para identificar los contribuyentes que están tanto en la base de datos de la DIAN como los que han declarado en las bases de datos en la Unidad de Rentas.

III. Resultados del análisis con la base de datos de la DIAN

La base de datos es muy homogénea, lo que facilita el análisis entre los formatos y los años; de esta manera, se procede a examinar las variables individualmente, tanto de las bases de datos de la DIAN como de la Unidad de Rentas.

Los resultados reportan una baja asimetría entre los impuestos y la base de datos de la DIAN (tabla 2).

Tabla 2. Empresas

Formato	Loterías	Degüello	Contribución
Formato 110	59	50	57
Formato 210	31	20	31
Formato 230	0	0	0
Formato 240	0	0	0
Concordancia	0	1	3

Fuente: elaboración propia.

Para el impuesto de loterías foráneas la DIAN reporta 59 empresa que no están en la base de datos de la Unidad de Rentas en el formato 110; en el formato 210 hay 31 empresas que reportan la actividad económica CIIU: 9.200 (actividades de juego y azar), y en el año 2018 no hay ninguna para los formatos 230 y 240.

Para el impuesto al degüello de ganado mayor se detectan 50 empresas en el formato 110, para el formato 210 hay 20 y solo una con simetría entre la base de datos de la DIAN y la Unidad de Rentas con actividad económica 1011.

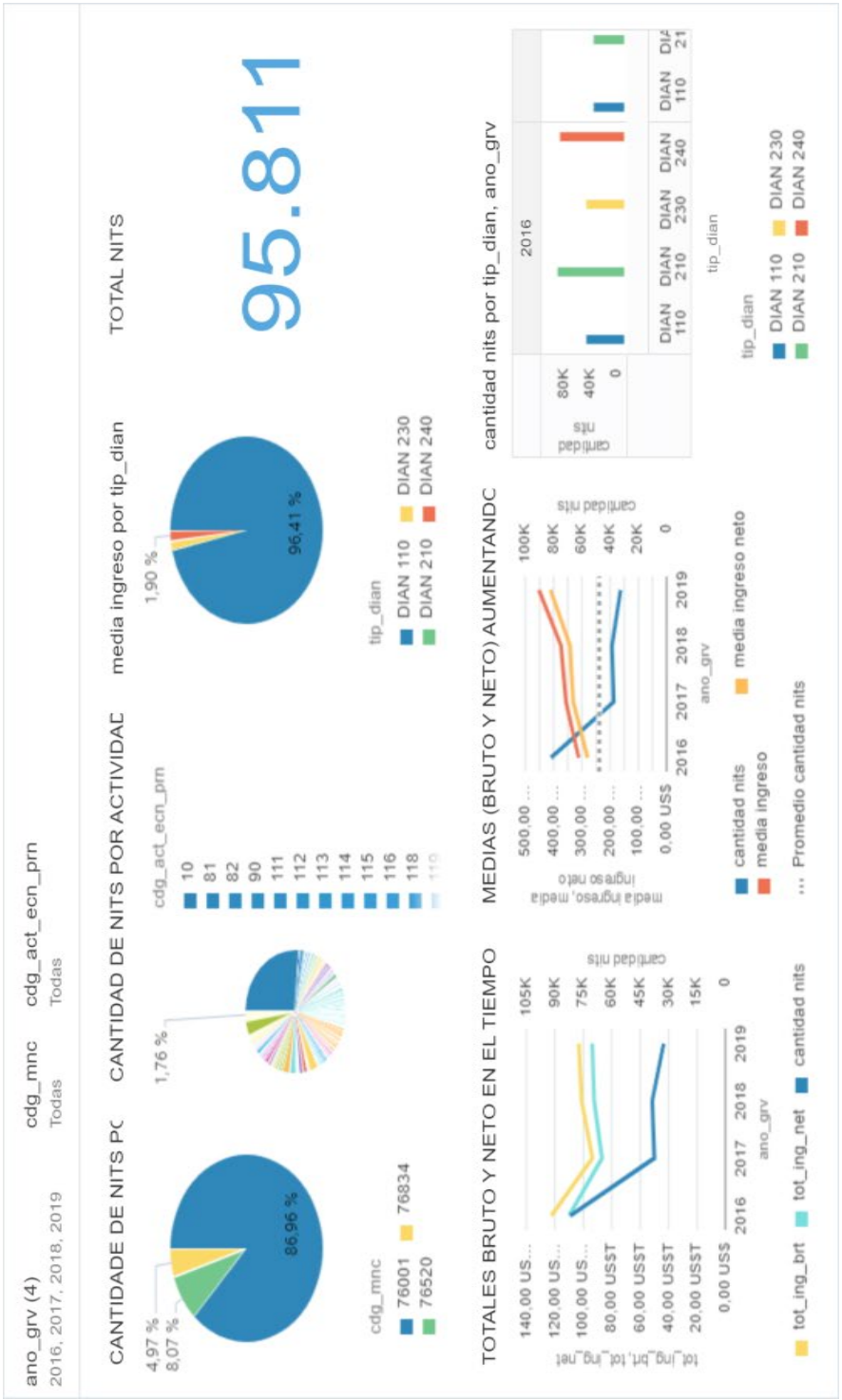
Para el impuesto de contribución para el deporte, la recreación y el aprovechamiento del tiempo libre el resultado es el siguiente: para el formato 110 se obtienen 57 empresas; para el formato 210, se detectan 31 empresas con las actividades 5512, 5530 y 5590, con solo 3 concordancias entre la base de datos de la DIAN y la Unidad de Rentas (figura 3).

IV. Análisis de la bases de datos de la Cámara de Comercio

Las bases de datos de las cámaras son poco homogéneas, cada cámara maneja un sistema diferente, sin una resolución de información exógena, lo que hace que su minería sea más árdua.

Para el análisis de las cámaras se realizan dos procedimientos, uno de ellos con el programa Python.

FIGURA 3. MODELO DE TABLERO DE MANDO INTEGRAL TRIBUTARIO



Fuente: Unidad Administrativa Especial de Rentas y Gestión Tributaria.

En la tabla 3 se detallan las cámaras de comercio a las que se les solicitó la información.

Tabla 3. Relación de Cámaras de Comercio del Valle del Cauca

Siglas	Nombre
CC Cali	Cámara de Comercio de Cali
CC Palmira	Cámara de Comercio de Palmira
CC Tuluá	Cámara de Comercio de Tuluá
CC Buenaventura	Cámara de Comercio de Buenaventura
CC Buga	Cámara de Comercio de Buga
CC Cartago	Cámara de Comercio de Cartago
CC Sevilla	Cámara de Comercio de Sevilla

Fuente: elaboración propia.

V. Procedimiento con el programa Python

Uno de los objetivos de esta investigación es la realización de modelos predictivos y prescriptivos para identificar los posibles evasores mediante las variables exógenas asociadas a cada contribuyente (tales como el tamaño de la empresa, antigüedad, fecha de renovación, ingresos, municipios, sector, barrio, entre otros). A través de los datos históricos el modelo es capaz de analizar datos de una persona jurídica, y con ellos predecir su situación tributaria actual. Asimismo, el modelo permite incorporar nuevos datos para fortalecer su capacidad de extrapolación (en este momento entrenados para degüello de ganado y contribuyentes al deporte)

Para hacer esto posible se utilizó el algoritmo XgBoost, parte de la familia de los gradientes ponderados. Este algoritmo, desarrollado en los últimos tres años, combina lo mejor del mundo de los gradientes y los árboles de decisión para la construcción de modelos predictivos. Este se ha labrado un lugar como un candidato usual en las competencias de Machine Learning como Kaggle o KDNuggets. El algoritmo permite, a partir de un número arbitrario de variables, seleccionar la combinación con mayor predictibilidad y extrapolarla a conjuntos de datos similares. El modelo compara combinaciones de variables por un número finito temporal mientras encuentra la forma de reducir al máximo el error.

Previo al modelo se ejecutan los siguientes procesos:

- Se realiza la limpieza y concatenado de la información X que alimenta al modelo. En este caso, la información exógena aprovisionada por las Cámaras de Comercio del Valle del Cauca.
- Estos datos son supremamente no balanceados en el sentido de que son muchísimos más aquellos omisos que los exactos e inexactos. Por tanto, para aprovechar las características del modelo se balancearon las categorías por predecir; es decir, siempre el modelo va a tender a predecir que un conjunto de contribuyentes está compuesto

de omisos, pero, al mismo tiempo, es especialmente sensible a las combinaciones de variables que permiten predecir las otras dos categorías.

- Para aprovechar la robustez del modelo, se ejecuta un proceso de ingeniería de variables. Es decir, se transforman las 79 variables de entrada por una serie de primitivos para construir casi 200 rasgos que alimentan al modelo. Este proceso es válido en cuanto queremos aumentar la capacidad predictiva sin aumentar el sesgo, no tenemos necesidad de explicar por qué las combinaciones funcionan de x o y manera, dado que esa es más una tarea de otras disciplinas (y métodos) como el análisis financiero forense y la econometría.
- Se realiza una división de validación cruzada de los datos en *splits* 80/10/10. Estas divisiones permiten comprobar el sesgo del modelo y prevenir que el mismo esté sobreajustado a los datos que sirvieron para el entrenamiento. Esto se cuantifica mediante el uso del área debajo de la Curva Precision-Recall. Se usa esta métrica, en vez del área bajo la curva tradicional, dado el no balance de los datos ya mencionado. La literatura en el tema sugiere esta corrección, que infortunadamente es menos gráfica, pero más correcta. Los distintos *splits* presentaron puntajes entre 0,91 a 0,96 (figura 4).

FIGURA 4. ESQUEMA DE ESTRUCTURA DEL PROGRAMA PYTHON

Fuente: Unidad Administrativa Especial de Rentas y Gestión Tributaria.

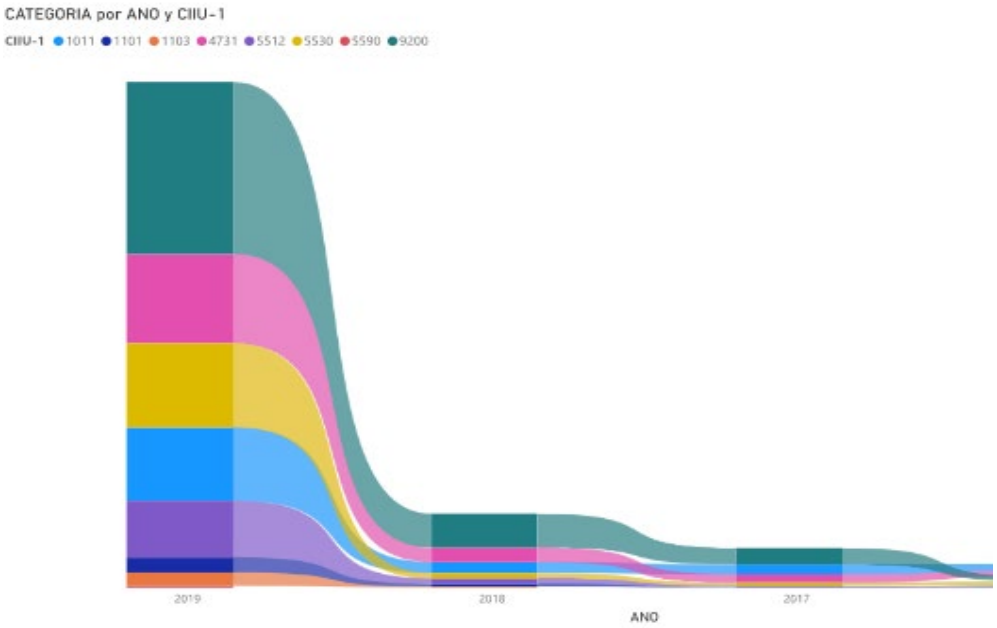
Con las cámaras de comercio se obtienen los siguientes resultados (tabla 4):

TABLA 4. RESULTADOS DEL PROGRAMA PYTHON

	Actividad económica	Degüelle	Loterías	Esparcimiento	Total general
Degüelle	1011	228			228
Esparcimiento	5512			178	178
Esparcimiento	5530			206	206
Esparcimiento	5590			4	4
Loterías	9200		559		559
	Total general	**228**	**559**	**388**	**1.175**

Fuente: elaboración propia

Figura 5. Resultados programa Python



Fuente: Cámara de Comercio.

VI. Conclusiones y recomendaciones

El método para la recopilación de programas de fiscalización es innovador, y se puede mejorar sustancialmente con la construcción de resoluciones de información exógena para la simetría de los datos entre las cámaras. Es una herramienta para la generación de programas de fiscalización a gran escala.

También se puede considerar un método que permita extrapolarse a otros impuestos y, de esta forma, desarrollar programas conjuntos entre entidades fiscalizadoras.

La renta de loterías no presenta una actividad económica específica como venta, en esta actividad se incluyen también los juegos de azar y venta de chance; se presentan muchas informaciones espurias en el resultado.

Se halló mayor efectividad con las bases de datos de las cámaras de comercio del departamento, por tanto, la utilización de un marco relacional entre la actividad económica y los impuestos es útil para la realización de programas de fiscalización. Aunque se requiere primero trabajar en los marcos legales e institucionales que permitan asegurar el cumplimiento de la ley y el respeto por la integridad de los contribuyentes.

Se recomienda un sistema de gestión de cambio en la Unidad de Rentas para acoger las nuevas herramientas y lograr su efectividad, fortaleciendo los procesos de recolección de datos con la estructuración de información exógena. Se deja la invitación para crear más estudios que incorporen algoritmos de asociación y árboles de decisión.

Es importante fortalecer y detallar los marcos legales e institucionales que permitan asegurar el cumplimiento de la ley, el respeto de las garantías individuales, la confidencialidad y el uso de la información oculta, así como los principios éticos de imparcialidad, seguridad y control de la gestión. También es importante invertir en capital humano, la capacitación es fundamental pues de esta depende la implementación exitosa y vertiginosa en temas de inteligencia fiscal tributaria, además de la actualización permanente de la tecnología con herramientas disruptivas y efectivas. También es importante contar con una voluntad política que impulse permanentemente la tecnología. En suma, es necesario fortalecer e invertir en ciberseguridad para garantizar la integridad de los sistemas y la seguridad de los datos personales.

Por último, los proveedores tecnológicos deben responder a las nuevas necesidades sobre calidad y fiabilidad de los datos provistos para estudios en inteligencia fiscal tributaria (caso SAR o sistemas de información que no poseen métricas completas para comparar bases de datos); también, realizar una solicitud ante la Federación Nacional de Departamentos para aumentar casillas en el formulario.

Referencias

Banco Interamericano de Desarrollo (BID) (2013). *Estado de la Administración Tributaria en América Latina: 2006-2010*. BID.

Comisión de Expertos para la Equidad y la Competitividad Tributaria (2015). Informe final presentado al ministro de Hacienda y Crédito Público. Fedesarrollo.

Normativa y jurisprudencia

Consejo de Estado, Sentencia 2005-01421 de 30 de octubre de 2013.

Estatuto Tributario Nacional.

Dirección de Impuestos y Aduanas Nacionales (DIAN) (2012). Resolución 000139 de 21 de noviembre.

Gobernación del Valle del Cauca (2017). Estatuto Tributario Departamental. Ordenanza No. 474.

Fecha de recepción: 30 de noviembre de 2020.
Aprobación par 1: 1 de febrero de 2021.
Aprobación par 2: 17 de febrero de 2021.