UNU WIDER

4
24

Tax research for development

Toolkit para la estimación de brechas tributarias mediante un enfoque ascendente

Mostafa Bahbah,[1] Sebastián Castillo,[2] Kwabena Adu-Ababio,[3] and Amina Ebrahim[3]

Diciembre 2024

WIDER TECHNICAL NOTE

wider.unu.edu

United Nations University
World Institute for Development Economics Research

Resumen: La presente nota se refiere al toolkit para la estimación de brechas tributarias, que incluye el código (archivos DO de Stata) y el archivo README (instrucciones para ejecutar el código). Esta nota describe la bibliografía y la metodología que fundamentaron el desarrollo del toolkit. El toolkit se refiere a la estimación de las brechas en el impuesto sobre el valor añadido (IVA), el impuesto sobre las sociedades y el impuesto sobre la renta individual mediante un enfoque ascendente, que utiliza aprendizaje automático para determinar los resultados de las auditorías operativas y los impuestos potenciales.

Palabras clave: toolkit, enfoque ascendente, auditorías operativas, aprendizaje automático

Códigos de clasificación JEL: H25, H26, H32

Agradecimientos: Los autores agradecen sinceramente los aportes de Jukka Pirttilä, Maria Jouste, Gerald Agaba y Hilja-Maria Takala. Los autores agradecen la financiación del International Tax Compact (ITC). El ITC facilita la Secretaría de la Addis Tax Initiative (ATI), que apoyó el desarrollo del conjunto de herramientas sobre la brecha fiscal. El ITC está financiado por el Ministerio Federal Alemán de Cooperación Económica y Desarrollo, cofinanciado por la Unión Europea, e implementado por la Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. Este trabajo forma parte del programa de Domestic Revenue Mobilization de UNU-WIDER, financiado por Norad.

Material complementario: El código y los archivos README relacionados con esta publicación pueden descargarse libremente desde la página web del toolkit.

Publicaciones relacionadas:

- Estimating tax gaps in Zambia: WIDER Working Paper 2023/25
- Estimating the value-added tax gap in Tanzania: WIDER Working Paper 2024/66

Esta nota técnica está disponible en inglés (original), francés y portugués.

[1] Universidad de Tampere, Finlandia; [2] Universidad de Helsinki, Finlandia, y Finnish Centre of Excellence in Tax Systems Research (FIT); [3] UNU-WIDER, Helsinki, Finlandia; correspondencia: amina@wider.unu.edu

Este estudio se ha elaborado en el marco del proyecto UNU-WIDER Tax research for development (phase 3), que forma parte del área de investigación Creating the fiscal space for development. El proyecto forma parte del programa de Domestic Revenue Mobilization, financiado mediante contribuciones específicas de la Agencia Noruega de Cooperación para el Desarrollo (Norad). El conjunto de herramientas sobre la brecha fiscal recibió apoyo financiero del International Tax Compact (ITC), que facilita la Secretaría de la Addis Tax Initiative (ATI). El ITC está financiado por el Ministerio Federal Alemán de Cooperación Económica y Desarrollo, cofinanciado por la Unión Europea, e implementado por la Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.

https://doi.org/10.35188/UNU-WIDER/WTN/2023-4

United Nations University World Institute for Development Economics Research – UNU-WIDER

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introducción

La movilización de ingresos internos para el financiamiento del gasto público es un aspecto fundamental para muchos gobiernos. En ese sentido, las autoridades fiscales desempeñan la función decisiva de diseñar sistemas tributarios eficientes y eficaces, que aseguren los máximos ingresos posibles con un mínimo de pérdidas. Sin embargo, los ingresos óptimos o máximos no logran alcanzarse ya que tanto los individuos como las empresas contribuyentes hacen todo lo posible para evitar o evadir sus obligaciones fiscales. Ante esta realidad, la investigación busca medir la brecha entre los ingresos fiscales reales y potenciales, buscando determinar a cuánto ascendería la pérdida de ingresos si todas las personas y empresas cumplieran plenamente con las normas tributarias.

Esta nota técnica ilustra el cálculo de las brechas tributarias, que se entienden como la diferencia entre los ingresos tributarios reales y los potenciales, basado en una metodología ascendente que hace uso de diversas formas de declaración y datos estadísticos. Además, esta nota acompaña el conjunto de herramientas (toolkit)sobre la brecha tributaria como una introducción a los conceptos generales sobre las brechas tributarias, con un énfasis en el enfoque ascendente utilizado para la estimación de estas. El toolkit desarrollado por UNU-WIDER busca simplificar los conceptos complejos de este enfoque de estimación, sirviendo como una guía para los usuarios sobre los métodos de medición sistemáticos que utilizan los parámetros disponibles y a su alcance.

Debido a la amplitud del concepto de brechas tributarias, existe una variedad de enfoques para su estimación. Sin embargo, todos se centran en investigar los motivos por los que los ingresos fiscales reales no corresponden a los ingresos potenciales. Algunos de los métodos generales utilizados habitualmente se enfocan en indicadores macroeconómicos agregados como punto de referencia para determinar la desviación, mientras que los métodos menos utilizados recurren a datos microeconómicos para estimar las brechas. La elección del método depende del acceso y la disponibilidad de los datos administrativos, lo que varía según las jurisdicciones. Si bien esta nota técnica presenta los principales enfoques para la estimación de las brechas tributarias, se centra en el uso de datos microeconómicos provenientes de auditorías operativas (en la mayoría de los casos) y en situaciones donde las auditorías aleatorias son raras.

En la presente nota, empezaremos por definir las brechas tributarias y veremos cómo evoluciona este concepto dependiendo de variables específicas de interés, relacionadas con las políticas y parámetros de cumplimiento, como se explica en la Sección 2. A continuación, en la Sección 3, abordaremos los métodos generales utilizados para la estimación de las brechas tributarias, entre los que destaca el enfoque ascendente que describiremos en sus diversas formas y métodos. En la Sección 4, describiremos los componentes del toolkit, que consta de dos etapas principales: la limpieza de datos y un método de aprendizaje automático (machine learning) para el cálculo de brechas tributarias.

2 Definición de la brecha tributaria

Las autoridades tributarias suelen encontrar diferencias considerables entre los ingresos fiscales esperados y los montos recaudados. Esta diferencia, denominada pérdida de recaudación, surge principalmente cuando los impuestos adeudados no se pagan dentro de un período determinado. Estos impuestos adeudados corresponden al monto de los impuestos que, teóricamente, podría recaudarse. Esto conduce al concepto de brecha tributaria, que se define como la diferencia entre

la recaudación real y la recaudación teórica de impuestos, es decir, si se respetara plenamente el código tributario.

Desde un punto de vista normativo, la brecha tributaria consta en general de dos componentes principales: la brecha de cumplimiento y la brecha normativa. La brecha de cumplimiento se refiere a la diferencia entre los ingresos reales obtenidos durante un año específico y los ingresos máximos que se podrían haber recaudado a partir de las actividades económicas que se llevaron a cabo durante dicho período. La brecha normativa se produce por decisiones legislativas que buscan modificar la reglamentación fiscal estándar mediante la incorporación de exenciones, deducciones o tasas reducidas para ciertos casos específicos (Hutton 2017). Los cambios en el marco normativo pueden aumentar o reducir la brecha normativa. Por ejemplo, si se aumenta el umbral para la exención de impuestos, lo que significa que un segmento más grande de los ingresos quedará exento de impuestos, o si se aplica una tasa impositiva reducida para un grupo específico de contribuyentes como las pequeñas empresas o las personas de bajos ingresos, la brecha normativa aumentaría, ya que la recaudación sería menor en comparación con la recaudación potencial máxima bajo las normas fiscales habituales. Por otra parte, la brecha normativa también podría aumentar sin alterar de forma alguna el marco normativo, debido a cambios en la composición de la base impositiva que impliquen que un segmento mayor del ingreso neto quede sujeto a la tasa impositiva estándar (Barra y otros 2023).

Por su parte, la brecha de cumplimiento se compone de dos elementos: la brecha de ajuste y la brecha de recaudación. La brecha de ajuste se genera principalmente por aquellas actividades económicas que las autoridades tributarias desconocen o no logran abarcar, incluyendo las actividades de entidades que no están registradas, no declaran, declaran menos o declaran erróneamente sus impuestos, tal como se observa en jurisdicciones con un alto nivel de informalidad. La brecha de recaudación se refiere a la diferencia entre las obligaciones fiscales calculadas, considerando las devoluciones y retenciones, y los impuestos que efectivamente se han pagado. Esto incluye los impuestos pendientes de pago de los cuales las autoridades fiscales tienen conocimiento pero que no han logrado cobrar, generalmente porque se encuentran paralizados por controversias tributarias, porque se considera que su cobranza supone un costo demasiado alto o porque son imposibles de cobrar mediante vías legales.

La bibliografía también identifica tres componentes distintos en la brecha de cumplimiento que complementan las brechas de ajuste y recaudación antes mencionadas (Gemmell and Hasseldine 2014; Durán-Cabré y otros 2019).

1. *Componente de subdeclaración:* los contribuyentes declaran ingresos inferiores a los que en realidad percibieron o solicitan más deducciones, créditos u otros beneficios fiscales de los permitidos por la ley, o una combinación de ambos. Esto genera una diferencia entre las obligaciones tributarias reales del contribuyente y el monto declarado.

2. *Componente de no declaración:* indica la brecha entre los declarantes potenciales y los que efectivamente declaran sus impuestos, lo que refleja el alcance de la no declaración de impuestos y la evasión fiscal.

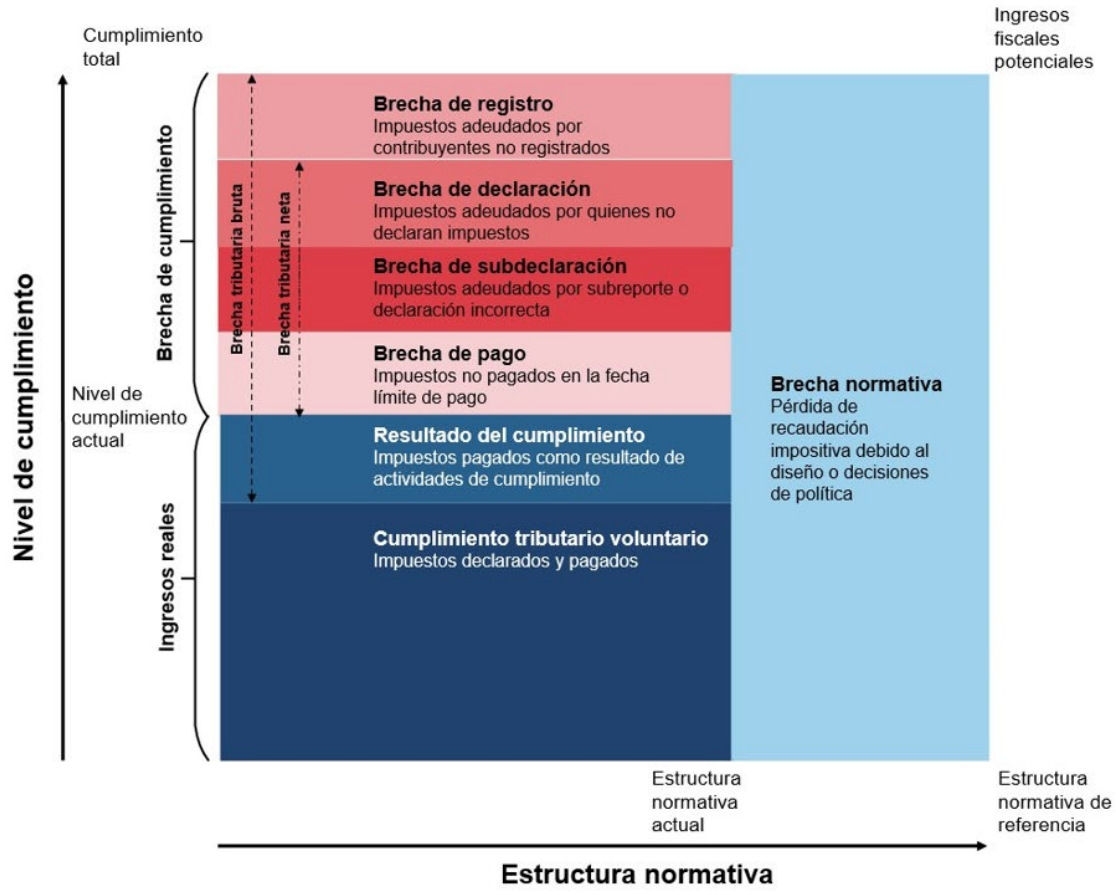
3. *Componente de incumplimiento de pago:* la diferencia entre los ingresos fiscales potenciales y los reales, que refleja la parte de los impuestos evadidos por no declarar o por subdeclarar montos a las autoridades fiscales.

4. *Componente de no registro:* Se refiere a la diferencia entre el número de entidades o personas físicas que deben registrarse para fines fiscales (como empresas, personas que trabajan por

cuenta propia o dueños de propiedades) y las que están realmente registradas. También se conoce como la brecha de registro.

Por último, desde el punto de vista de la recaudación, algunas autoridades fiscales dividen la brecha tributaria en dos categorías: la brecha tributaria bruta y la brecha tributaria neta.[1] Por ejemplo, el Servicio de Impuestos Internos de los Estados Unidos (IRS, por sus siglas en inglés) define la brecha tributaria bruta como la diferencia entre las obligaciones tributarias totales reales exigidas por la ley para un año fiscal específico y el monto de los impuestos que los contribuyentes pagan voluntaria y oportunamente para ese año. Por otro lado, la brecha tributaria neta se refiere al monto restante adeudado del total de las obligaciones tributarias tras contabilizar todos los pagos obtenidos mediante medidas de fiscalización, así como los pagos atrasados realizados voluntariamente para un año fiscal determinado (Plumley 2005). El gráfico 1 ilustra los componentes clave de la brecha tributaria total y la superposición entre las distintas definiciones de sus componentes.

Gráfico 1: Conceptos de la brecha tributaria



Nota: ilustración simplificada de los conceptos de la brecha tributaria.

Fuente: ilustración de los autores.

[1] Cabe destacar que las definiciones de la brecha tributaria bruta y neta pueden variar levemente entre las distintas autoridades tributarias, lo cual es un reflejo de los entornos de fiscalización tributaria y las prioridades administrativas de cada país.

3 Metodologías para determinar la brecha tributaria

Existen dos enfoques generales para la estimación de la brecha tributaria: el enfoque descendente y el enfoque ascendente. El enfoque descendente utiliza datos agregados como indicadores macroeconómicos o datos de cuentas nacionales para evaluar todas las pérdidas fiscales de manera integral, calculando la brecha como la diferencia entre los ingresos potenciales estimados y los ingresos reales. Sin embargo, este enfoque no permite identificar los orígenes de la brecha tributaria ni explicar por qué determinadas áreas o actividades quedan exentas de impuestos. En cambio, el enfoque ascendente se basa en datos microeconómicos provenientes de las administraciones tributarias, incluyendo los resultados de las auditorías aleatorias u operativas enfocadas en criterios específicos, así como en otros datos administrativos generales de las autoridades tributarias para evaluar el grado de incumplimiento en segmentos específicos del sistema tributario, grupos determinados de contribuyentes o tipos de incumplimiento (Hutton 2017).

3.1 Ventajas y desventajas del enfoque ascendente

Ventajas

El enfoque ascendente para la estimación de las brechas tributarias ofrece varias ventajas respecto a otras metodologías, especialmente por su capacidad de proporcionar conocimientos detallados (estimaciones granulares) basados en auditorías fiscales. Sus principales ventajas son:

Una mayor precisión mediante el uso de datos detallados: El método ascendente aprovecha datos granulares provenientes de auditorías financieras, lo que permite obtener estimaciones más precisas de la brecha tributaria. Esta técnica contrasta con las estrategias descendentes, que se basan en indicadores económicos más amplios y por lo tanto podrían omitir sutilezas en los comportamientos de contribuyentes individuales o en sectores específicos.

Conocimientos detallados que permiten acciones precisas: Las estrategias ascendentes ofrecen una comprensión detallada del cumplimiento tributario a nivel individual o empresarial. Este nivel de detalle permite que las autoridades fiscales diseñen intervenciones precisas para sectores, categorías de contribuyentes o casos de incumplimiento específicos, mejorando así la eficiencia y el impacto de las medidas para asegurar el cumplimiento (Hutton 2017).

Permite abordar el sesgo de selección en la estimación de la brecha tributaria: El sesgo de selección constituye un obstáculo importante para el cálculo preciso de la brecha tributaria, debido a la naturaleza no representativa de los contribuyentes seleccionados para las auditorías fiscales. La aplicación del enfoque ascendente, especialmente en combinación con técnicas de aprendizaje automático, es una forma eficaz de reducir este sesgo. En efecto, este método no se basa en suposiciones sobre la distribución de los datos, lo que brinda protección ante cualquier tipo de sesgo que pudiera distorsionar la estimación de la brecha tributaria (Alaimo Di Loro y otros 2023).

Análisis sectorial: El enfoque ascendente permite un análisis pormenorizado de las brechas de cumplimiento tributario, sector por sector. Gracias a estos conocimientos detallados, las autoridades tributarias pueden orientar sus estrategias de cumplimiento de forma más específica, enfocándose en los sectores que presentan las mayores brechas. Esta estrategia focalizada puede traducirse en una mayor eficiencia en la recaudación de impuestos sin necesidad de aumentar las actividades de fiscalización y control de manera general (Barra y otros 2023; Hutton 2017).

Adaptabilidad a distintos tipos de impuestos: La flexibilidad del enfoque ascendente permite utilizarlo para calcular las brechas de distintos tipos de impuestos, incluyendo el impuesto sobre el valor

añadido (IVA), el impuesto sobre las sociedades y el impuesto sobre la renta individual. Esta adaptabilidad es crucial dado que cada tipo de impuesto tiene sus propios desafíos de cumplimiento y tácticas de evasión fiscal.

Mayor cumplimiento fiscal: Un enfoque ascendente puede brindar una visión del comportamiento de los contribuyentes, lo que a su vez permite verificar o perfeccionar los modelos existentes para identificar y gestionar los riesgos. También contribuye a identificar errores específicos que podrían manejarse de forma más efectiva recurriendo a métodos alternativos, como educar a los contribuyentes, mejorar los servicios o llevar a cabo más auditorías y reevaluaciones (Barra y otros 2023).

Permite establecer límites máximos y mínimos para las estimaciones: El enfoque ascendente permite la aplicación de múltiples técnicas a la misma unidad de contribuyentes, además de permitir el análisis estadístico de sensibilidad de los resultados (Barra y otros 2023).

Desventajas

A pesar de los beneficios del enfoque ascendente, la bibliografía también pone de relieve las siguientes limitaciones (Warren 2018; Fiscalis Tax Gap Project Group 2018).

Endogeneidad: Este método se basa principalmente en el conocimiento y los datos existentes dentro de la administración tributaria, lo que reduce su capacidad para identificar factores desconocidos o elementos que habían pasado inadvertidos.

Dificultades para considerar elementos desconocidos: Dado que el método se basa en datos conocidos y resultados de operaciones, tiene dificultades para considerar factores que no son fáciles de observar, como los ingresos subdeclarados. Tampoco abarca la economía sumergida, ya que generalmente para las auditorías solo se selecciona a contribuyentes ya registrados. Como resultado, las estimaciones de estos elementos desconocidos a menudo implican ajustes aproximados, lo que puede reducir su exactitud.

Perspectiva limitada: Este enfoque va de lo particular a lo general, centrándose en los contribuyentes de manera individual. Si bien esto proporciona conocimientos detallados, puede pasar por alto tendencias o patrones macroeconómicos.

Riesgo de agregación: Los enfoques ascendentes solamente estiman los componentes de la brecha tributaria, lo que obliga a realizar una agregación para determinar la brecha total. Sin embargo, este proceso conlleva un riesgo de doble cómputo y de sobreestimación de la brecha tributaria total, especialmente cuando existen superposiciones entre distintos componentes de la brecha.

3.2 Tipos de auditorías

Las autoridades fiscales suelen basarse en la información de auditorías para predecir la evasión fiscal y estimar las brechas tributarias. Estas auditorías pueden dividirse en dos categorías principales: las auditorías aleatorias y las auditorías operativas. Ambas tienen objetivos distintos y metodologías propias que ofrecen perspectivas sobre el cumplimiento de los contribuyentes.

Auditorías aleatorias

Las auditorías aleatorias consisten en elegir muestras de contribuyentes mediante un proceso aleatorio con el objetivo de reflejar con la mayor exactitud posible la población general que se busca representar. En estas auditorías, todos los contribuyentes seleccionados se someten a una revisión minuciosa para identificar cualquier discrepancia entre los montos que declararon en sus

impuestos y los montos que están obligados a declarar por ley. Los resultados de estas auditorías ofrecen un indicador confiable del nivel de cumplimiento dentro del grupo de muestra. Para poder extrapolar los resultados de la muestra a toda la población, es necesario garantizar que el proceso de selección sea completamente aleatorio y no incluya ningún tipo de criterios de selección (Barra y otros 2023).

A pesar de su exhaustividad, las auditorías aleatorias también presentan desventajas según Feinstein (1999), incluyendo los costos elevados tanto para las autoridades fiscales como para los contribuyentes, especialmente aquellos que acatan las normativas tributarias. También existe un lapso entre el período que cubren los datos y el momento en que los resultados están disponibles. Los rendimientos financieros suelen ser más bajos que los de las auditorías selectivas, ya que se evalúa tanto a contribuyentes que cumplen como a contribuyentes que no, a diferencia de las auditorías selectivas que se enfocan en aquellos con mayor probabilidad de evadir impuestos. Además, no permiten identificar a los contribuyentes no inscritos, lo que se traduce en subestimaciones de algunas brechas tributarias.

Por último, las autoridades fiscales podrían mostrarse reticentes a llevar a cabo auditorías aleatorias debido a la percepción pública de la autoridad por parte de los contribuyentes. Los contribuyentes que sí cumplen podrían considerar las auditorías aleatorias como una intromisión excesiva o una fiscalización injusta, generando una opinión pública negativa y mermando la confianza en las autoridades fiscales.

Auditorías operativas

Las auditorías operativas se basan en la evaluación de riesgos y se enfocan en contribuyentes específicos seleccionados según criterios determinados por el análisis de riesgos de las autoridades fiscales. Estas auditorías pueden enfocarse en uno o varios tipos de impuestos y abarcar todo el espectro de cada uno de ellos o solamente un segmento específico. Debido a los criterios de selección, este tipo de auditoría puede no ser representativo de toda la población, ya que no todos los contribuyentes tienen la misma posibilidad de ser seleccionados como sucede con las auditorías aleatorias. Por lo tanto, las autoridades fiscales implementan una estimación ascendente de la brecha tributaria utilizando datos de auditorías no aleatorias, recurriendo a métodos que buscan deducir las características de la población general a partir de una muestra no representativa.

3.3 Procedimientos de estimación ascendente

Se pueden emplear diversos métodos para aplicar un enfoque ascendente. Todos se basan en información de auditorías para predecir el comportamiento de las empresas o los períodos no auditados. En esta sección, examinaremos las estimaciones más comunes y destacaremos sus principales características (ventajas y desventajas).

Técnicas de regresión

Las técnicas de regresión son habituales en los estudios sobre el enfoque ascendente y pueden corregir el sesgo de selección. También pueden ayudar a determinar características que permiten predecir si un contribuyente cumplirá sus obligaciones y estimar el grado de incumplimiento. Entre estas técnicas de regresión se encuentran el método de Heckman y el de propensity score matching.

El método de Heckman. El método de Heckman aborda el sesgo de selección, que se produce durante el proceso de las auditorías operativas y que genera endogeneidad en el subconjunto de los contribuyentes auditados. Este método, que se basa en el trabajo de Heckman (1979), implica un proceso de estimación en dos etapas. La primera etapa determina la probabilidad de que una

observación se incluya en la muestra, lo que básicamente significa calcular la probabilidad de que un contribuyente sea seleccionado para una auditoría, mediante una ecuación de regresión probit. La segunda etapa se centra en estimar la variable de interés, que en este caso corresponde al monto recuperado tras la auditoría. Para ello se consideran las variables explicativas y un regresor específico que corrige el sesgo de selección. Este regresor específico, conocido como la razón inversa de Mills, se deriva de los parámetros estimados en la ecuación de selección. A continuación, la ecuación de resultado se calcula mediante la regresión lineal de mínimos cuadrados ordinarios, incorporando un factor de la ecuación de la primera etapa.

El Fiscalis Tax Gap Project Group (2018) señala que existen dos aspectos importantes a tener en cuenta al estimar la brecha tributaria utilizando el método de Heckman. En primer lugar, la ecuación de selección debe ser capaz de explicar los resultados eficazmente, dado que el método depende en gran medida de la capacidad de la ecuación para predecir el incumplimiento. En segundo lugar, la ecuación debe incorporar al menos una variable que influya en la selección para la auditoría, pero que no afecte directamente el incumplimiento. Esto contribuye a evitar estimaciones erróneas debido a la multicolinealidad. Básicamente, para una estimación precisa de la brecha tributaria, es necesario tener datos sobre los factores que conducen a la auditoría de dichos contribuyentes y que no estén directamente relacionados con el nivel de incumplimiento. En la práctica, esta restricción de exclusión es difícil de cumplir.

Método de propensity score matching. El método de propensity score matching se utiliza para corregir el sesgo de selección mediante la ponderación de los datos. Este método empieza con el cálculo de una «puntuación de la propensión» para cada entidad, utilizando modelos estadísticos para estimar su probabilidad de incumplimiento o de auditoría. Un modelo de selección binaria determina la propensión mediante probit o logit. Una vez estimadas estas puntuaciones, el método hace corresponder las entidades auditadas con las que no lo han sido pero que comparten puntuaciones de propensión similares. Para establecer las correspondencias se puede utilizar alguno de los siguientes métodos: vecino más cercano, *caliper*, *kernel* o local lineal. Una vez establecidas las correspondencias, el paso final consiste en asignar un valor a las declaraciones no auditadas. Este valor, conocido como N, es un valor imputado o estimado de lo que indicaría la declaración si se hubiera sometido a auditoría. Esta imputación se basa en los valores reales observados en las declaraciones auditadas correspondientes. Este paso es necesario para estimar cuál habría sido el nivel de cumplimiento tributario del grupo no auditado si se hubiese sometido a una auditoría.

Método de agrupación (clustering)

Este enfoque clasifica a los contribuyentes auditados y no auditados en grupos basados en variables relevantes utilizadas para seleccionar la empresa a auditar, como el tamaño de la empresa, la región geográfica y el sector industrial. Permite calcular la brecha tributaria total mediante la suma de las brechas estimadas para cada grupo. Estas estimaciones se obtienen aplicando un factor de escala a los resultados de las auditorías de los contribuyentes auditados, proyectando estos hallazgos a la población general de cada grupo. Pese a ser fácil de aplicar e implementar, este método solo corrige el sesgo de selección de forma parcial, produciendo resultados que no son completamente fiables.

Análisis de valores extremos

El análisis de valores extremos aprovecha el sesgo de selección hacia los contribuyentes con los mayores niveles de incumplimiento en las auditorías operativas. Se ocupa del comportamiento de los valores extremos (máximos o mínimos) en un conjunto de datos en vez de los valores promedio, entendiendo que, sin importar la distribución general de los datos, los valores extremos a menudo siguen una distribución de Pareto generalizada. Esto sugiere que se pueden obtener

conocimientos sobre la tasa general de incumplimiento tributario de las grandes empresas a partir de un número reducido de casos extremos (es decir, los mayores evasores de impuestos). Este enfoque es pertinente cuando los datos presentan características de la distribución de Pareto, una forma de distribución de probabilidades que sostiene que un pequeño porcentaje de los casos contribuye de forma desmedida al valor total observado en los datos, como cuando la subdeclaración de impuestos es extremadamente asimétrica (con unas pocas grandes empresas que son responsables de la mayor parte de la brecha) (Bloomquist y otros 2014).

Enfoques de aprendizaje automático

Si bien la aplicación de métodos de aprendizaje automático en estudios económicos es relativamente reciente, es un campo que se está abriendo progresivamente, en particular en temas relacionados con los impuestos, como la evasión fiscal, el fraude y la predicción del cumplimiento, así como para mejorar las auditorías fiscales y la estimación de las brechas tributarias. Aunque la investigación en el ámbito tributario generalmente se basa en métodos tradicionales para el cálculo de predicciones, estas técnicas presentan limitaciones debido a su dependencia de los métodos de regresión lineal y los estrictos supuestos de distribución que implican. En la práctica, los datos suelen presentar patrones más complejos, por lo que estos métodos no son lo suficientemente flexibles para realizar predicciones. Por ello, algunos estudios han empezado a recurrir a métodos de aprendizaje automático para mejorar los resultados de las predicciones.

A modo de ejemplo de la aplicación del aprendizaje automático, Pérez López y otros (2019) utilizaron un modelo de red neuronal de perceptrón multicapa para predecir el fraude fiscal a partir de datos completos de las declaraciones del impuesto sobre la renta individual en España. Este método de aprendizaje automático logró predecir tanto la probabilidad como la verosimilitud de cada contribuyente de cometer prácticas fraudulentas. Zumaya y otros (2021) utilizaron dos algoritmos de aprendizaje automático, incluyendo redes neuronales profundas y bosques aleatorios para predecir la evasión del IVA en México mediante el análisis de los datos transaccionales y las redes de interacción de los contribuyentes. El estudio reveló que la combinación de estos tres métodos permitía identificar nuevos sospechosos potenciales al aprender de los patrones de los evasores ya conocidos. Ioana-Florina y Mare (2021) intentaron predecir la propensión de los contribuyentes a evadir impuestos según su confianza en el sistema tributario, utilizando un modelo de red neuronal de perceptrón multicapa. Este enfoque mostró mejores resultados predictivos, superando los del modelo de regresión logística binaria.[2]

Por otra parte, los métodos de aprendizaje automático también se pueden utilizar para contribuir a los esfuerzos de auditoría fiscal. Por ejemplo, Howard y otros (2020) evaluaron el potencial de las técnicas de aprendizaje automático para optimizar el proceso de selección de casos para las auditorías por correspondencia realizadas por el Servicio de Impuestos Internos de los Estados Unidos (IRS, por sus siglas en inglés). El estudio reveló que, en algunas categorías de auditoría, los métodos de aprendizaje automático superan a las técnicas tradicionales en la clasificación y selección de declaraciones de impuestos para las auditorías por correspondencia. Estos métodos no solamente generan mayores ingresos, sino que también reducen la tasa de auditorías sin cambios, lo que significa que hay menos auditorías que no conducen a ninguna corrección en comparación con otros métodos. De la misma forma, Battaglini y otros (2022) utilizaron datos administrativos de las autoridades tributarias italianas para explorar el potencial de las técnicas de

[2] Véase también Alsadhan (2023); Baghdasaryan y otros (2022); Holtzblatt y Engler (2022); Murorunkwere y otros (2022, 2023); Raikov (2021); Savic´ y otros (2022) para otros ejemplos de la aplicación de métodos de aprendizaje automático para la predicción del fraude y la evasión de impuestos.

aprendizaje automático, como los bosques aleatorios, para mejorar la detección de la evasión fiscal y la consiguiente recaudación al optimizar el proceso de selección de los contribuyentes para las auditorías. El estudio indica que, en algunos casos, el aprendizaje automático podría mejorar la detección de evasiones fiscales hasta en un 83 por ciento y lograr una recaudación de hasta un 65 por ciento de los montos correspondientes.

La investigación sobre la estimación de las brechas fiscales no se quedaba atrás de estos recientes avances. Ante las limitaciones de los enfoques para la estimación de brechas tributarias antes mencionados, que se basan en métodos de regresión tradicionales para la elaboración de predicciones, algunos investigadores y autoridades fiscales empezaron a incorporar el uso de técnicas semiparamétricas a los métodos tradicionales, así como a utilizar el aprendizaje automático para mejorar los resultados de las predicciones. Si bien el aprendizaje automático supera a los métodos tradicionales en materia de predicción, también resulta eficaz para corregir el sesgo de selección en las estimaciones de las brechas tributarias basadas en auditorías operativas.

Para abordar el problema del sesgo de selección en el contexto de la estimación de las brechas tributarias, es importante distinguir los dos grandes tipos de sesgo de selección: el sesgo de selección causal y el sesgo de selección muestral. El sesgo de selección causal afecta la estimación de parámetros insesgados en el análisis causal, tal como ocurre cuando no se asignan los grupos de control y tratamiento de manera aleatoria, lo que conduce a estimaciones sesgadas sobre los efectos del tratamiento. Pero el que nos interesa es el sesgo de selección muestral, que se produce cuando la muestra utilizada para el entrenamiento de un modelo predictivo es distinta de la muestra utilizada para la predicción. En el caso de la estimación de brechas tributarias basada en auditorías operativas, este sesgo surge porque la muestra utilizada para el entrenamiento está compuesta por contribuyentes auditados seleccionados según ciertos criterios conocidos por las autoridades tributarias y que no son representativos de toda la población de contribuyentes, mientras que la muestra para la predicción incluye contribuyentes no auditados. De no abordarse adecuadamente, esta discrepancia puede resultar en predicciones sesgadas.

Un aspecto crucial en el manejo del sesgo de selección muestral es saber distinguir los sesgos generados por factores observables de los generados por factores no observables. El sesgo de selección observable se produce cuando el proceso de selección, por ejemplo, la decisión de auditar se basa en variables conocidas y cuantificables. En tales casos, si la probabilidad de auditoría puede calcularse certeramente a partir de estas covariables observables, se puede corregir el sesgo incluyendo estas covariables en el modelo de aprendizaje automático. Esta metodología concuerda con las estrategias descritas por Brewer y Carlson (2024), quienes recomiendan controlar el sesgo de selección mediante la incorporación de los factores observables. Se puede mitigar el sesgo de selección calculando e integrando la probabilidad de selección en el modelo, asumiendo que las decisiones de auditoría se fundamentan principalmente en datos observables.[3]

En situaciones donde el proceso de selección se rige por factores no observables que no están reflejados en el conjunto de datos, aumenta la complejidad del sesgo. Los métodos tradicionales pueden no ser suficientes para contrarrestar este tipo de sesgo. Estos casos requieren recurrir a técnicas más avanzadas, como incorporar una función de control basada en el método de Heckman en el modelo de aprendizaje automático para abordar el sesgo de selección basado en factores no

[3] Cabe presumir que las autoridades fiscales disponen de información sobre cómo decidir a quién auditar. Esta información suele ser confidencial, pero se puede aprovechar en el modelo de aprendizaje automático para predecir resultados con precisión. Sugerimos no incluir la relevancia de las covariables en la predicción, puesto que esta información pertenece al proceso de auditoría. No obstante, esos resultados también se pueden utilizar para mejorar el proceso de toma de decisiones en las auditorías.

observables (Brewer y Carlson 2024). En investigaciones recientes, encontramos ejemplos destacados de la incorporación de enfoques de aprendizaje automático en métodos tradicionales, así como estudios que calculan las brechas tributarias utilizando principalmente técnicas de aprendizaje automático.

Alaimo Di Loro y otros (2023) propusieron un método basado en aprendizaje automático que consiste en aplicar el algoritmo de potenciación del gradiente en dos etapas. Este método aborda el sesgo de selección generado por el uso de datos de auditorías no aleatorias y ofrece predicciones exactas. En primer lugar, el método estima las puntuaciones de propensión de que un contribuyente sea auditado utilizando un modelo de clasificación basado en la potenciación del gradiente, con árboles de clasificación y regresión (CART, por sus siglas en inglés) como aprendices base. Para ello, se dividen los datos en conjuntos de entrenamiento y de prueba. Luego, durante el proceso de entrenamiento, se seleccionan las covariables importantes. El resultado de este paso son las probabilidades estimadas de auditoría de cada empresa según sus covariables.

En segundo lugar, el método utiliza un modelo de regresión con potenciación de gradiente aplicando CART como aprendices base para predecir la base imponible potencial, incluyendo el IVA no declarado, y por ende los montos evadidos por cada empresa. En esta etapa, se utilizan las puntuaciones de propensión obtenidas anteriormente para crear ponderaciones para cada contribuyente, corrigiendo así cualquier representación excesiva o insuficiente en la muestra auditada. Al comparar este enfoque de aprendizaje automático con el modelo tradicional de Heckman, el aprendizaje automático resulta netamente superior para captar la variabilidad en la base imponible potencial y proporcionar predicciones más acertadas de la brecha tributaria.

Adu-Ababio y otros (2024) utilizaron algoritmos de aprendizaje automático supervisado con información de declaraciones de impuestos y auditorías para estimar las brechas tributarias en Zambia. La red neuronal artificial (artificial neural network) fue el principal algoritmo utilizado en este estudio, con una implementación en dos etapas. En la primera etapa, se recurrió únicamente a los datos de auditorías para crear de forma aleatoria iteraciones de múltiples versiones de conjuntos de datos de entrenamiento y de prueba, utilizando el 90 por ciento de los datos para entrenar el modelo. Analizando distintos parámetros tributarios, a cada iteración el algoritmo aprende del conjunto de datos de entrenamiento. Luego, el algoritmo utiliza lo que ha aprendido para predecir las tasas de evasión fiscal a partir de los datos de prueba. A continuación, el algoritmo compara las tasas de evasión fiscal calculadas con las reales. Si las predicciones no corresponden a la realidad, se incorporan mejoras al modelo y se repite el proceso hasta lograr resultados satisfactorios. En la segunda etapa, se despliega el modelo utilizando la muestra completa, con los datos auditados en el conjunto de entrenamiento y los datos no auditados en el conjunto de prueba. Una vez que el modelo ha aprendido de las variables explicativas seleccionadas, calcula la evasión fiscal a partir de los datos de prueba y luego utiliza tanto la evasión calculada como la evasión real para estimar las brechas tributarias. Los autores también utilizaron otros algoritmos de aprendizaje automático, como el bosque aleatorio (random forest), para comprobar la estabilidad y fiabilidad del método principal, obteniendo resultados bastante similares.

En la misma línea, Ebrahim y otros (2024) utilizaron datos tributarios y de auditorías en su estudio para estimar la brecha del IVA en Tanzania mediante aprendizaje automático, utilizando un algoritmo de bosque aleatorio. El objetivo era predecir los montos de la evasión fiscal de empresas auditadas y no auditadas en períodos en donde no se realizaron auditorías. Los autores compararon los resultados del método de aprendizaje automático con los de la regresión tradicional de mínimos cuadrados ordinarios. Observaron una reducción notable de la raíz del error cuadrático medio (RMSE, por sus siglas en inglés) junto con un incremento de los valores del coeficiente de determinación al utilizar el algoritmo de bosque aleatorio, lo que indica que este último logra predicciones más precisas. Los resultados indican una brecha del IVA promedio de alrededor del

62 por ciento, con variaciones considerables entre distintos sectores económicos. El sector agrícola en particular presentó la mayor brecha del IVA.

Otros avances en técnicas de aprendizaje automático incluyen el uso de métodos de regresión convencionales. Chudý y otros (2020) aplicaron una selección de muestras semiparamétrica al modelo de Heckman para estimar la brecha del impuesto sobre las sociedades en Eslovaquia. Esta extensión del modelo de Heckman supera al tradicional ya que permite un supuesto de realidad más flexible y un mejor modelado de las estructuras de datos complejas, así como el manejo de las relaciones no lineales y la heterocedasticidad propias de los datos. En la primera etapa de este modelo, se calculó la ecuación de selección mediante un método no paramétrico, como el suavizado de *kernel* (*kernel smoothing*), lo que permitió obtener aproximaciones flexibles de las distribuciones. Luego, en la segunda etapa del modelo, la ecuación de resultado incorporó estas estimaciones de la primera etapa para lograr una corrección más efectiva del sesgo de selección y capturar las relaciones más complejas que podrían ser omitidas por un modelo de regresión lineal. El estudio concluyó que este enfoque tuvo un mejor desempeño que otros enfoques alternativos, como el método de propensity score matching y la regresión lineal de mínimos cuadrados ponderados, tanto para abordar el sesgo de selección como para ofrecer predicciones más acertadas.

Las autoridades fiscales también empezaron a recurrir a técnicas de aprendizaje automático para mejorar sus estimaciones de las brechas tributarias o sus procesos de auditoría. La agencia tributaria italiana (s. f.) utilizó una combinación de aprendizaje automático con otros métodos tradicionales para calcular la brecha del IVA mediante lo que denominaron un método asistido por aprendizaje automático. La primera etapa de este método busca abordar el sesgo de selección propio de las auditorías no aleatorias mediante regresión logística, segmentando la población en grupos que tengan una probabilidad similar de ser auditados. A continuación, la población se estratifica en quintiles a partir de estas probabilidades, con lo cual los contribuyentes auditados son representativos de la población total de cada grupo. En la segunda etapa, se utiliza el aprendizaje automático, en este caso árboles de regresión por agregación de bootstrap (bagging), para realizar predicciones dentro de cada estrato. La última etapa busca mejorar la precisión de las predicciones utilizando el modelo de correspondencia basada en la media predicha (predictive mean matching o PMM), que se basa en las predicciones iniciales para asociar cada contribuyente no auditado (el receptor) con un contribuyente auditado (el donante) según la similitud de sus valores estimados. Este proceso asegura que los valores imputados reflejen fielmente la distribución de la variable objetivo, lo que permite inferir con precisión diversas características de distribución más allá de los simples promedios.

La agencia tributaria de Canadá (2019) utiliza una técnica de aprendizaje automático no supervisado para identificar grupos dentro de una población, siguiendo la misma lógica que describimos anteriormente en el primer paso de la metodología adoptada en Italia, donde los elementos de cada grupo se parecen más entre sí que a los de los demás grupos. Este algoritmo de aprendizaje automático clasifica automáticamente a las empresas en grupos basados en características específicas, dando por supuesto que las empresas no auditadas dentro de cada grupo comparten el mismo nivel de incumplimiento en relación con sus ingresos brutos declarados que las empresas auditadas. Se utilizó este método para obtener una estimación máxima y se combinó con el análisis de valores extremos para obtener una estimación de límite inferior de la brecha tributaria.

Resumen final

Existen diversos métodos para la estimación de brechas tributarias mediante enfoques ascendentes. Sin embargo, dependiendo del contexto y de los datos utilizados, algunos podrían ser

más adecuados que otros. De forma general, se puede aplicar un enfoque ascendente a partir de datos de auditorías aleatorias o basadas en riesgos. Para muchos investigadores, los datos provenientes de auditorías aleatorias son ideales para la estimación ascendente de brechas tributarias. Sin embargo, en muchos casos, las autoridades fiscales prefieren llevar a cabo auditorías basadas en riesgos, lo que plantea algunos desafíos para la estimación: los contribuyentes seleccionados para la auditoría podrían ser muy distintos de los demás contribuyentes, con lo que los resultados de la auditoría no serán representativos de la población general que no cumple sus obligaciones tributarias. En este caso, una estimación tradicional de mínimos cuadrados ordinarios puede no ser la mejor elección debido al sesgo de selección del proceso de auditoría. Por ende, los investigadores recurren a otros métodos para obtener estimaciones imparciales. A continuación, resumiremos las ideas clave sobre los métodos mencionados en esta sección.

Aunque el método de Heckman de dos etapas es uno de los más utilizados para corregir el sesgo de selección, a veces resulta difícil cumplir su restricción de exclusión, lo que podría provocar un aumento de los errores estándar debido a la multicolinealidad, y además tiende a subestimar la brecha fiscal ya que a menudo se omite la evasión fiscal y el incumplimiento no detectado. El método de propensity score matching ayuda a eliminar el sesgo de selección al crear grupos asociados de contribuyentes que cumplen y de contribuyentes que no cumplen basándose en características observables, lo que permite atribuir de forma más exacta las diferencias en los resultados de cumplimiento tributario al incumplimiento más que a factores no observados. Algunas autoridades fiscales aplican el enfoque de agrupamiento para detectar comportamientos inusuales y subdeclaraciones de impuestos dentro de grupos específicos, y luego estiman la brecha tributaria de cada grupo extrapolando los resultados de auditoría de los contribuyentes auditados a toda la población de ese grupo específico. Por otro lado, en comparación con los demás métodos, el análisis de valores extremos es más directo y eficiente en términos de tiempo y recursos. Sin embargo, requiere más supuestos, especialmente para la configuración de la cola en la distribución de Pareto, en la cual se basa para el modelado.

A diferencia de los métodos anteriores, las técnicas de estimación basadas en aprendizaje automático presentan importantes ventajas, especialmente en la gestión de relaciones complejas y no lineales, así como de factores no observados que influyen en el sesgo de selección. Se puede optar por métodos de aprendizaje automático por la flexibilidad que ofrecen y sus capacidades de predicción superiores.

4 El toolkit

En esta sección explicaremos los componentes del toolkit. El objetivo del toolkit es estimar la brecha tributaria del impuesto sobre el valor añadido (IVA), el impuesto sobre las sociedades y el impuesto sobre la renta individual. El toolkit está conformado por dos elementos principales: la limpieza de datos y la estimación. El proceso de limpieza de datos asegura la armonización y coherencia de los archivos de datos necesarios para la estimación ascendente. Además, permite alinearlos con los requisitos generales para la estimación mediante aprendizaje automático. Este es un punto importante considerando que los datos provienen de distintas fuentes y períodos, por lo que su estandarización facilita el proceso de estimación. La estimación se basa en la metodología de los bosques aleatorios, una técnica de aprendizaje automático. Ofrecemos una explicación técnica de esta metodología en el apéndice A.

4.1 Limpieza de datos

El proceso de limpieza de datos puede dividirse en tres etapas principales: las dos primeras se ocupan de los datos administrativos de declaración de impuestos (el IVA, el impuesto sobre las sociedades y el impuesto sobre la renta individual) y de auditoría, y la última muestra cómo combinar estos archivos de datos para su posterior análisis. Este paso tiene por finalidad procesar las distintas fuentes de datos, armonizarlas, y crear una estructura única que reúna información sobre los contribuyentes, las declaraciones de impuestos y los resultados de las auditorías y evaluaciones.

Generalmente, la información sobre las declaraciones de impuestos (IVA, impuesto sobre las sociedades e impuesto sobre la renta individual) se encuentra en archivos distintos a los de las auditorías, puesto que estas últimas se llevan a cabo después de que los individuos o empresas presentan sus respectivas declaraciones. Sin embargo, puede ser que en las declaraciones de impuestos se encuentren dos trámites para un mismo contribuyente. Esto puede ocurrir porque el contribuyente actualiza su declaración en algún punto dentro o fuera del plazo de declaración. Este es un problema de duplicación frecuente que se presenta en las bases de datos de las autoridades tributarias. En estos casos, un mismo dato aparece dos veces para el mismo contribuyente. En otras palabras, para un mismo contribuyente en un año fiscal específico, existen dos o más copias de la misma información. Uno de los principales objetivos de la etapa de limpieza de datos es asegurarse de que cada contribuyente se distinga de forma única por sus datos de identificación y el año de presentación de la declaración. En la primera etapa del toolkit, presentamos escenarios que pueden generar estos errores de duplicación y demostramos cómo el usuario puede tratarlos de manera individual. Es importante resolver todos los duplicados en todos los datos requeridos, tanto de declaración de impuestos como de auditoría, ya sea que se presenten en un archivo único o en varios archivos. En el caso de encontrarse en múltiples archivos, la estrategia consiste comenzar por resolver los duplicados y luego reunir los conjuntos de datos respectivos en un solo archivo.

En este punto del toolkit, también abordamos problemas en cuanto a los períodos de auditoría detectados en los datos de auditoría y su relación con períodos de declaración específicos. En algunos casos, los datos de auditoría están identificados por el año de evaluación, que también es el año de la declaración. A veces, es más bien el año de auditoría el que se utiliza como año de declaración. Sea cual sea el caso, es necesario identificar el año específico en los datos de auditoría que corresponde al año de declaración y combinarlos para obtener un solo archivo si la información está dispersa en varios archivos. Esto asegura que cada evaluación de auditoría esté debidamente vinculada a un período de declaración específico.

Al finalizar estas dos primeras etapas, consolidamos los datos de declaración de impuestos a nivel anual. Al calcular las brechas tributarias de forma anual, también nos aseguramos de que las evaluaciones de auditoría estén vinculadas a los resultados anuales, incluso si las auditorías se llevaron a cabo para múltiples años de declaración. Generalmente, esta consolidación se aplica al IVA y al impuesto sobre la renta individual, pero no al impuesto sobre las sociedades ya que siempre se declara de manera anual. Este procedimiento garantiza que contemos con una declaración de impuestos o resultado de auditoría (en caso de que el contribuyente haya sido auditado) por contribuyente por año.

Finalmente, combinamos los archivos de datos necesarios teniendo en cuenta que las variables de las declaraciones de impuestos y de las auditorías se encuentran en dos archivos distintos. Es fundamental comprender el proceso de combinación, ya que demuestra cuán eficaces hemos sido en limpiar y eliminar los duplicados en todos los archivos de datos. El objetivo es combinar la información correspondiente a la misma unidad (contribuyente) para el mismo período (año-mes).

Además, queremos que la información proporcionada por los datos de auditoría, como el resultado de la auditoría para un año de declaración específico, esté integrada a la del registro de impuestos en el período de declaración correspondiente. Por ejemplo, combinamos el registro de impuestos correspondiente al año 2018 con el resultado de la auditoría de datos inexactos de ese mismo año, si la empresa fue sometida a auditoría. No se dispondrá de dicha información si la empresa no fue auditada. Este es uno de los problemas frecuentes que encontrará el usuario, dado que las auditorías se realizan de manera retroactiva y para un número limitado de contribuyentes en base a declaraciones anteriores. Explicaremos cómo obtener los resultados de auditoría para estas empresas no auditadas en la próxima etapa del toolkit.

4.2 Estimación mediante aprendizaje automático

Adoptamos un enfoque ascendente para calcular la brecha tributaria. Para su aplicación se requiere el resultado de la auditoría del contribuyente, que asumimos como indicador de la declaración de datos inexactos. Esta variable se obtiene durante el proceso de auditoría y se encuentra en los datos de auditoría al término del mismo. Sin embargo, solo las empresas que han sido auditadas cuentan efectivamente con esta variable. Por esta razón, es necesario estimar o predecir los resultados para los contribuyentes y períodos no auditados, ya que la información proveniente de auditorías se limita a períodos y unidades específicos. Así, un contribuyente que fue auditado en el año de declaración 3 no fue auditado en el año de declaración 2, lo que significa que debemos incorporar una predicción para los períodos no auditados para garantizar que disponemos de toda la información requerida. Se requiere un procedimiento de estimación para calcular predicciones acertadas sobre la declaración de datos inexactos.

En el toolkit, aplicamos el método del bosque aleatorio para predecir la declaración de datos inexactos en las empresas y los períodos no auditados. Esta metodología permite calcular estimaciones detalladas ya que identifica mejor los posibles valores atípicos que el método de predicción lineal. Es necesario calibrar el bosque aleatorio definiendo dos parámetros fundamentales: la cantidad de iteraciones (o árboles) y la cantidad de usos que predecir en cada división. Para ello, es necesario usar datos que contengan la variable a predecir en el caso de la declaración de datos inexactos. Por ende, en primer lugar, el conjunto de datos se divide entre datos auditados y no auditados. El primer conjunto se utilizará para afinar el modelo y el segundo para realizar las predicciones.

Es necesario dividir los datos de auditoría en muestras de entrenamiento y de prueba para el proceso de calibración. El objetivo es mejorar la precisión de las estimaciones, puesto que la metodología utiliza los datos de entrenamiento para analizar las variables y luego compara la predicción con el valor real en los datos de prueba. Este proceso permite obtener los dos parámetros esenciales. Por otro lado, los parámetros aseguran que el error de predicción, en otras palabras, la diferencia entre la predicción y el valor real, sea lo más bajo posible. Utilizando el modelo óptimo, el toolkit realiza una comparación con un modelo de regresión para demostrar la exactitud de la predicción y así validar el modelo de predicción.

Finalmente, se calcula la brecha tributaria. En primer lugar, el modelo se aplica solamente a los datos auditados, dado que dichas observaciones contienen información sobre la declaración de datos inexactos. En este paso, el modelo determina el índice (o ponderación) de cada variable auxiliar (o covariante). Luego, el modelo genera predicciones de los datos no auditados utilizando el índice óptimo y se obtienen las predicciones de declaración de datos inexactos. La brecha tributaria se calcula sumando la variable de declaración de datos inexactos (predicha o descubierta mediante auditoría) a la declaración de impuestos, obteniendo así el monto potencial de impuestos. La brecha tributaria corresponde a la relación entre la declaración de datos inexactos y el monto potencial del impuesto e indica el porcentaje del monto potencial no recaudado a causa de la

declaración inexacta. Esta variable se obtiene a partir del tipo de grupo (por ejemplo, el sector) y demuestra el nivel de detalle de la metodología.

5 Observaciones finales

En este toolkit, buscamos desarrollar un marco de trabajo práctico para la estimación de brechas tributarias en el impuesto sobre el valor añadido (IVA), el impuesto sobre las sociedades y el impuesto sobre la renta individual, aplicando una metodología ascendente. Este toolkit ha sido diseñado para que las autoridades y los responsables de las políticas tributarias puedan calcular la diferencia entre los ingresos fiscales reales recaudados y los ingresos potenciales que podrían haberse recaudado si se respetara plenamente la normativa tributaria. Ofrece un marco estandarizado que se puede aplicar en países en desarrollo, considerando su contexto y los recursos a su disposición. Nuestro enfoque se basa en la aplicación de un algoritmo de aprendizaje automático que utiliza datos de declaraciones de impuestos simplificadas y de auditorías para predecir la declaración de datos inexactos y el incumplimiento y luego estimar las brechas tributarias tanto a nivel agregado como por sectores o regiones específicos.

En esta nota, analizamos las definiciones de las brechas tributarias para brindar una comprensión de sus componentes, puesto que el objetivo de nuestro método es estimar las brechas generadas por la subdeclaración, la declaración de datos inexactos y el incumplimiento entre los contribuyentes inscritos. A continuación, examinamos los procedimientos tradicionales comúnmente utilizados, destacando los beneficios de recurrir a la estimación mediante aprendizaje automático con un enfoque ascendente respecto a otras técnicas de estimación.

El toolkit puede dividirse en dos etapas fundamentales: la gestión de datos y el análisis mediante aprendizaje automático. Durante la etapa de gestión de datos, se preparan los conjuntos de datos tributarios y de auditorías para su posterior análisis, mediante un proceso de limpieza, gestión de duplicados y combinación de datos para asegurar su armonización y facilitar su transición hacia las etapas de aprendizaje automático. El aprendizaje automático predice la declaración de datos inexactos para los contribuyentes y períodos que no han sido objeto de auditorías mediante la aplicación de algoritmos de bosque aleatorio. Al entrenarse a partir de datos de auditorías, estos modelos tienen la capacidad de proporcionar estimaciones adecuadas de la evasión fiscal de los casos no auditados, permitiendo un cálculo integral de la brecha tributaria. Asimismo, este toolkit ofreció una comparación de los resultados de los modelos de regresión tradicionales con los modelos de aprendizaje automático para poner de relieve su mayor capacidad de predicción.

Por último, ofrecemos algunas sugerencias de proyectos futuros que expandan y mejoren el toolkit actual. Se podría recurrir a otros algoritmos de aprendizaje automático, como las redes neuronales, y comparar la exactitud de las predicciones entre diversos métodos. Otro aspecto para considerar es la adaptación del toolkit a otros lenguajes de programación además de Stata para ampliar su alcance. Asimismo, se requiere más investigación en cuanto a las formas de implementar el toolkit en diversos contextos nacionales. Por último, el toolkit puede ofrecer un punto de partida para investigaciones futuras sobre el comportamiento de los contribuyentes, con el fin de brindar asistencia a las autoridades en el diseño y la implementación de mejores medidas de control y estrategias que aseguren el cumplimiento.

Bibliografía

Adu-Ababio, K., Koivisto, A., and Mwale, E. (2024). *Estimating tax gaps in Zambia.* (Preprint)

Alaimo Di Loro, P., Scacciatelli, D., and Tagliaferri, G. (2023). '2-step Gradient Boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy'. *Statistical Methods & Applications*, 32(1): 237–270. https://doi.org/10.1007/s10260-022-00643-4

Alsadhan, N. (2023). 'A Multi-Module Machine Learning Approach to Detect Tax Fraud'. *Computer Systems Science and Engineering*, 46(1): 241–253. https://doi.org/10.32604/csse.2023.033375

Athey, S., and Imbens, G. W. (2019). 'Machine learning methods that economists should know about'. *Annual Review of Economics*, 11(1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Baghdasaryan, V., Davtyan, H., Sarikyan, A., and Navasardyan, Z. (2022). 'Improving tax audit efficiency using machine learning: The role of taxpayer's network data in fraud detection'. *Applied Artificial Intelligence*, 36(1): 2012002. https://doi.org/10.1080/08839514.2021.2012002

Barra, P. A., Hutton, M. E., and Prokof'yeva, P. (2023). *Corporate Income Tax Gap Estimation by using Bottom-Up Techniques in Selected Countries: Revenue Administration Gap Analysis Program.* Washington, DC: International Monetary Fund. https://doi.org/10.5089/9798400246265.005

Battaglini, M., Guiso, L., Lacava, C., Miller, D. L., and Patacchini, E. (2024). 'Refining public policies with machine learning: The case of tax auditing'. *Journal of Econometrics*, 105847. https://doi.org/10.1016/j.jeconom.2024.105847

Békés, G., and Kézdi, G. (2021). 'Regression Trees'. In *Data analysis for business, economics, and policy* (pp. 417–437). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108591102.015

Bloomquist, K. M., Hamilton, S., and Pope, J. (2014). 'Estimating Corporation Income Tax Under-Reporting Using Extreme Values from Operational Audit Data'. *Fiscal Studies*, 35(4): 401–419. https://doi.org/10.1111/j.1475-5890.2014.12036.x

Brewer, D., and Carlson, A. (2024). 'Addressing sample selection bias for machine learning methods'. *Journal of Applied Econometrics*, 39(3): 383–400. https://doi.org/10.1002/jae.3029

Canada Revenue Agency (2019). *Tax gap and compliance results for the federal corporate income tax system.*

Chudý, M., Gábik, R., Bukovina, J., and Šrámková, L. (2020). *Searching for gaps: Bottom-up approach for Slovakia.* Institute for Financial Policy (IFP).

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). 'Random forests'. In C. Zhang and Y. Ma (eds), *Ensemble ma- chine learning: Methods and applications* (pp. 157–175). New York: Springer. ttps://doi.org/10.1007/978- 1-4419-9326-7_5

Durán-Cabré, J. M., Esteller Moré, A., Mas-Montserrat, M., and Salvadori, L. (2019). 'The tax gap as a public management instrument: application to wealth taxes'. *Applied Economic Analysis*, 27(81): 207–225. https://doi.org/10.1108/AEA-09-2019-0028

Ebrahim, A., Castillo, S., Leyaro, V., Swema, E., and Haule, O. (2024). *Estimating the Value-Added Tax Gap for SMMEs in Tanzania: An Empirical Analysis.* (Manuscript)

Feinstein, J. S. (1999). 'Approaches for estimating noncompliance: examples from federal taxation in the United States'. *The Economic Journal*, 109(456): 360–369. https://doi.org/10.1111/1468-0297.00439

FISCALIS Tax Gap Project Group (2018). 'The Concept of Tax Gaps: Corporate Income Tax Gap Estimation Methodologies'. Working paper 73 – 2018. Luxembourg: Publications Office of the European Union. (European Commission's Directorate-General Taxation and Customs Union) https://doi.org/10.2778/83206

Gemmell, N., and Hasseldine, J. (2014). 'Taxpayers' behavioural responses and measures of tax compliance 'gaps': A critique and a new measure'. *Fiscal Studies*, 35(3): 275–296. https://doi.org/10.1111/j.1475- 5890.2014.12031.x

Hartshorn, S. (2016). *Machine learning with random forests and decision trees: A Visual guide for beginners*. Kindle edition.

Heckman, J. J. (1979). 'Sample selection bias as a specification error'. *Econometrica*, 47(1): 153–161. https://doi.org/10.2307/1912352

Holtzblatt, J., and Engler, A. (2022). *Machine Learning and Tax Enforcement*. Tax Policy Center, Urban Institute & Brookings Institution.

Howard, B., Lykke, L., Pinski, D., and Plumley, A. (2020). 'Can Machine Learning Improve Correspondence Audit Case Selection? Considerations for Algorithm Selection, Validation, and Experimentation'. In A. Plumley (ed.), *The IRS Research Bulletin: Proceedings of the 2020 IRS / TPC Research Conference* (pp. 147–169). Internal Revenue Service.

Hutton, M. E. (2017). *The Revenue Administration–Gap Analysis Program: Model and Methodology for Value-Added Tax Gap Estimation*. International Monetary Fund. https://doi.org/10.5089/9781475583618.005

Ioana-Florina, C., and Mare, C. (2021). 'The utility of neural model in predicting tax avoidance behavior'. In I. Czarnowski, R. Howlett, and L. Jain (eds), *Intelligent Decision Technologies: Proceedings of the 13th KES-IDT 2021 Conference* (pp. 71–81). https://doi.org/10.1007/978-981-16-2765-1_6

Italian Revenue Agency (n.d.). *Italy: VAT gap estimation via bottom up approach*.

Murorunkwere, B. F., Haughton, D., Nzabanita, J., Kipkogei, F., and Kabano, I. (2023). 'Predicting tax fraud using supervised machine learning approach'. *African Journal of Science, Technology, Innovation and Development*, 15(6): 731–742. https://doi.org/10.1080/20421338.2023.2187930

Murorunkwere, B. F., Tuyishimire, O., Haughton, D., and Nzabanita, J. (2022). 'Fraud detection using neural networks: A case study of income tax'. *Future Internet*, 14(6): 168. https://doi.org/10.3390/fi14060168

Pérez López, C., Delgado Rodríguez, M. J., and de Lucas Santos, S. (2019). 'Tax fraud detection through neural networks: An application using a sample of personal income taxpayers'. *Future Internet*, 11(4): 86. https://doi.org/10.3390/fi11040086

Plumley, A. (2005). 'Preliminary update of the tax year 2001 individual income tax underreporting gap estimates'. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* (Vol. 98, pp. 19–25).

Raikov, A. (2021). 'Decreasing tax evasion by artificial intelligence'. *IFAC-PapersOnLine*, 54(13): 172–177.

Savic´, M., Atanasijevic´, J., Jakovetic´, D., and Krejic´, N. (2022). 'Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method'. *Expert Systems with Applications*, 193(May): 116409. https://doi.org/10.1016/j.eswa.2021.116409

Schonlau, M., and Zou, R. Y. (2020). 'The random forest algorithm for statistical learning'. *The Stata Journal*, 20(1): 3–29. https://doi.org/10.1177/1536867X20909688

Varian, H. R. (2014). 'Big data: New tricks for econometrics'. *Journal of Economic Perspectives*, 28(2): 3–28.

Warren, N. (2018, April). 'Estimating Tax Gap is Everything to an Informed Response to the Digital Era'. In *13th International Revenue Administration Conference on Tax System Integrity in a Digital Age* (p. 1-41). Disponible en la dirección: https://ssrn.com/abstract=3200838 (última revisión: 23 de junio de 2019)

Zacharis, N. Z. (2018). 'Classification and regression trees (CART) for predictive modeling in blended learning'. *IJ Intelligent Systems and Applications*, 3(1): 9. https://doi.org/10.5815/ijisa.2018.03.01

Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G., and Pineda, C. (2021). 'Identi- fying tax evasion in Mexico with tools from network science and machine learning'. In O. Granados and J. Nicolás-Carlock (eds), *Corruption networks: Concepts and applications* (pp. 89–113). Cham: Springer. https://doi.org/10.1007/978-3-030-81484-7_6

A APÉNDICE - Algoritmo de bosque aleatorio

El bosque aleatorio es uno de los algoritmos de aprendizaje automático por conjuntos más utilizados y con mejor desempeño para tareas de predicción (Athey e Imbens 2019).[4] A diferencia de los modelos de regresión tradicionales, que presuponen relaciones lineales y tienen dificultades cuando el número de observaciones es inferior a las variables independientes, el bosque aleatorio puede gestionar relaciones no lineales en los datos y evita el problema de estimar más parámetros de los que los puntos de datos pueden admitir. Además, detecta mejor la existencia de valores atípicos, produciendo predicciones más exactas en dichos casos (Athey e Imbens 2019). Esto se debe a que no utiliza todas las variables predictoras al mismo tiempo, obteniendo así predicciones más acertadas que los métodos de regresión tradicionales (Schonlau y Zou 2020). Además de ser muy sencillo de utilizar, el bosque aleatorio es fácil de comprender y de rápida implementación. Además, posee un buen desempeño en comparación con otros algoritmos de aprendizaje automático (Varian 2014).

Un bosque aleatorio nos permite predecir la variable objetivo (y) a partir de variables de entrada (x). Básicamente, consiste en una colección de árboles de decisión creados utilizando subconjuntos aleatorios de datos. Pero ¿qué son los árboles de decisión y cómo se utilizan para crear un modelo de bosque aleatorio? Para responder esta pregunta, empezaremos por explicar los conceptos de los árboles de decisión y su funcionamiento y luego describiremos cómo crear un modelo de bosque aleatorio y utilizarlo para ejecutar tareas de predicción.

A1 Árboles de decisión

Los árboles de decisión son un tipo de algoritmo de aprendizaje automático supervisado y se utilizan tanto para tareas de regresión como de clasificación. Lo que hacen es dividir los datos en subconjuntos basados en los valores de las características de las variables de entrada (x) para predecir valores (y). Este proceso de división continúa hasta que los datos dentro de cada subconjunto sean tan homogéneos como sea posible respecto a la variable objetivo. Este método también se conoce como el algoritmo de árboles de clasificación y regresión (CART, por sus siglas en inglés), que es un sistema para encontrar la mejor división en cada paso para optimizar la exactitud de las predicciones.

Algoritmo CART

Tipos de CART:

- Los árboles de clasificación son un tipo de algoritmo de árboles de decisión utilizado para clasificar variables objetivo categóricas. Consisten en segmentar el espacio de predicción en distintas regiones, donde cada región corresponde a una etiqueta de clase específica. El objetivo es determinar a qué categoría pertenece la variable objetivo según los atributos de entrada.

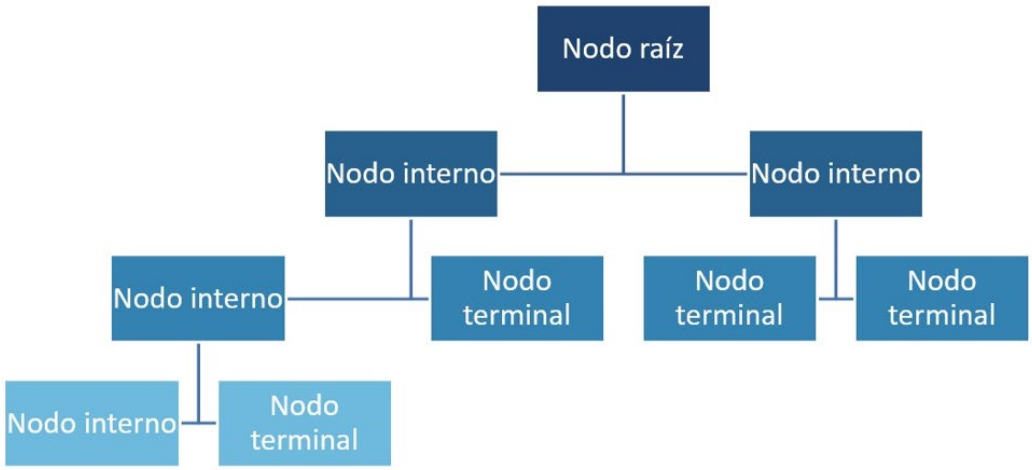
- Los árboles de regresión son un tipo de algoritmo de árbol de decisión diseñado para predecir variables objetivo continuas. Dividen el espacio de predicción en regiones y proporcionan como resultado un valor continuo para cada región.

[4] Los métodos por conjuntos combinan varios modelos simples, conocidos como aprendices débiles, para crear un único modelo predictivo más sólido.

¿Cómo funciona el algoritmo CART?

La creación de un árbol de decisión mediante el método CART comienza en el nodo raíz, que representa la totalidad del conjunto de datos. Este nodo raíz es el punto de partida del árbol. A partir de allí, el algoritmo identifica el mejor atributo para dividir el conjunto de datos y asigna dicho atributo al nodo. Así, se van formando ramas que conducen a nodos internos: cada noto interno representa una decisión basada en el valor del atributo elegido. Los datos se siguen dividiendo en cada nodo interno, generando más ramas y nodos. Este proceso se repite, creando una estructura jerárquica. Los extremos de estas ramas son los nodos terminales, que son los que proporcionan la predicción final como se muestra en el gráfico 1. En las tareas de clasificación, la predicción en un nodo terminal corresponde a la clase predominante de las observaciones en dicho nodo, y en las tareas de regresión corresponde al valor promedio de las observaciones.

Gráfico A1: Estructura de un árbol de decisión



Fuente: ilustración de los autores.

En las tareas de regresión, el método CART utiliza la reducción residual como criterio de división. Esto significa que los datos se dividen en cada nodo para minimizar la diferencia cuadrática media entre los valores predichos y los reales, con el objetivo de lograr el error residual más bajo posible. Para las tareas de clasificación, el algoritmo CART utiliza la impureza de Gini para evaluar todas las divisiones potenciales, optando por aquella que reduzca la impureza de forma más efectiva, aumentando así la pureza de los subconjuntos resultantes. La impureza de Gini mide la probabilidad de clasificar incorrectamente una instancia aleatoria según la clase predominante en un subconjunto. Este proceso de división es recursivo y continúa hasta que se cumplan ciertos criterios de parada. Estos criterios incluyen: llegar a un nodo donde todos los registros compartan el mismo valor objetivo, que el tamaño del nodo sea inferior a un umbral definido por el usuario, que el árbol alcance su profundidad máxima predefinida, que un nodo tenga menos de un número definido de casos, o que una división no mejore la pureza de manera significativa (Zacharis 2018).

Un riesgo fundamental que se presenta al utilizar árboles de decisión es el sobreajuste del modelo. Esto puede pasar si el modelo crece sin limitaciones, como cuando un árbol de regresión sigue dividiéndose hasta que queda una sola observación en cada nodo terminal. Aunque este puede ser un resultado casi perfecto para los datos de entrenamiento (véase la definición en la Nota más adelante), afecta la capacidad de generalización del modelo a nuevos datos desconocidos. Generalmente, los modelos sobreajustados tienen un buen desempeño con los datos de entrenamiento, pero funcionan muy mal con los datos de validación o de prueba, ya que han interiorizado más el ruido que la información relevante.

Para abordar los problemas de sobreajuste, el método CART utiliza una técnica de poda una vez que el árbol se ha generado por completo. La poda implica recortar el árbol para eliminar los nodos que aportan un valor predictivo mínimo, simplificando así el modelo y mejorando su capacidad de generalización. Una técnica muy utilizada es la poda de complejidad de costo, que implica, en primer lugar, generar un árbol grande a partir de un parámetro de complejidad muy pequeño para asegurar que se evalúen todas las divisiones potenciales. A continuación, se eliminan las divisiones de manera secuencial y se reevalúa el desempeño del modelo mediante validación cruzada. Este proceso continúa hasta que las podas no aporten más a la optimización del modelo (Békés y Kézdi 2021).

El gráfico A2 presenta un ejemplo de pseudocódigo para un algoritmo de generación de árboles para tareas de clasificación y regresión. Supongamos un escenario en el que queremos crear un árbol de decisión para predecir una variable objetivo utilizando un conjunto de datos X que contiene múltiples covariables A y la variable objetivo y. El parámetro «tarea» indica si se trata de una clasificación o una regresión.

El algoritmo empieza por generar un único árbol T con un nodo raíz. Si se cumplen todos los criterios de parada, el algoritmo procede a etiquetar el nodo. Para las tareas de clasificación, se etiqueta el nodo con la clase predominante entre las muestras en X. Para las tareas de regresión, se etiqueta el nodo con el valor medio de y.

Si no se cumplen los criterios de parada, el algoritmo busca el mejor atributo $a \in A$ que divide el conjunto de datos X de la manera más efectiva. Las tareas de clasificación se llevan a cabo aplicando una función de impureza, como la impureza de Gini. Para las tareas de regresión, el algoritmo busca reducir al mínimo la varianza dentro de los nodos. Entonces se etiqueta el nodo con el atributo a.

Gráfico A2: Pseudocódigo de generación de árboles para tareas de clasificación y regresión

Algoritmo 1 Algoritmo de generación de árboles `growingtree`(*X , A, y,* task)

Requiere: Conjunto de datos para entrenamiento *X*, conjunto de atributos *A*, variable resultante *y*, tarea (clasificación o regresión)

Garantizar: Árbol de decisión

1: Empieza con un único árbol T con un nodo superior

2: Si se cumplen todos los criterios de detención

3: si task == clasificación entonces

4: *T* tiene un nodo con la clase más común en *X* como etiqueta

5: else

6: *T* tiene un nodo con la media de *y* en *X* como etiqueta

7: end if

8: else

9: encuentre $a \in A$, que mejor divide *X* usando la función de impuridad (para clasificación) o minimiza la varianza (para regresión)

10: Etiquete el nodo con *a*

11: para un posible valor *v* de *a* haga

12: X_v = el subconjunto de *X* que tiene $a = v$

13: $A_v = A - a$

14: `growingtree`(X_v, A_v, *y,* task)

15: conecte el nuevo nodo con el nodo superior con etiqueta *v*

16: end for

17: end if

18: return `pruningtree`(*X , A, y,* task)

Nota

En aprendizaje automático, dividimos los datos en dos subconjuntos principales:

Conjunto de entrenamiento: Este subconjunto se utiliza para construir modelos como árboles de regresión y bosques aleatorios. Incluye características de entrada (variables independientes) y la variable objetivo (variable dependiente). El modelo aprende patrones y relaciones a partir de estos datos.

Conjunto de prueba: Este subconjunto se utiliza para evaluar el desempeño del modelo. Durante la etapa de entrenamiento, el modelo no tiene acceso al conjunto de prueba, lo que permite una evaluación imparcial de la capacidad de generalización del modelo a datos nuevos y desconocidos.

A continuación, el algoritmo recorre todos los valores posibles *v* del atributo elegido *a*. Para cada valor *v*, crea un subconjunto *X* donde el atributo *a* asume el valor *v*. También actualiza el conjunto de atributos *A* eliminando el atributo *a*. A continuación, el algoritmo se ejecuta de manera recursiva para seguir desarrollando el árbol, utilizando el subconjunto de *X* y el conjunto actualizado de atributos *A*. Este proceso recursivo continúa, conectando nuevos nodos al nodo raíz con etiquetas que corresponden a los valores *v*.

Una vez que el árbol ha alcanzado su máximo desarrollo de acuerdo con los criterios iniciales, el algoritmo procede a podar el árbol. El proceso de poda se lleva a cabo mediante una función de poda independiente que evalúa si la eliminación de ciertos nodos y ramas mejora el desempeño del árbol en un conjunto de datos de prueba. Para ello, se utilizan técnicas de validación cruzada con el fin de asegurar la capacidad de generalización del árbol a datos desconocidos.

Mediante la repetición de este proceso, el algoritmo de generación de árboles crea un árbol de decisión que divide el conjunto de datos X en regiones cada vez más reducidas. Cada nodo terminal (hoja) del árbol corresponde a una región específica dentro del espacio de características. En las tareas de clasificación, el nodo terminal representa la clase predominante dentro de esta región, mientras que, en las tareas de regresión, representa el valor promedio de y.

A2 Bosque aleatorio (random forest)

Pese a su utilidad, los árboles de decisión presentan limitaciones importantes, en particular su tendencia a sobreajustar los datos incluso tras el proceso de poda. En situaciones reales, es común que los datos sean desordenados y contengan anomalías que no se generalizan adecuadamente. Los árboles de decisión pueden crear divisiones muy específicas que corresponden bien a los datos de prueba, pero que no dan resultados acertados en datos nuevos y desconocidos. Los bosques aleatorios responden a este problema utilizando múltiples árboles de decisión y haciendo un promedio de sus resultados. Generar varios árboles a partir del mismo conjunto de datos no resuelve el problema, dado que se obtendrían árboles similares. En su lugar, los bosques aleatorios generan árboles utilizando subconjuntos aleatorios de los datos. El uso de subconjuntos variados garantiza que los árboles sean diferentes, lo que contribuye a nivelar las anomalías y mejorar la precisión de las predicciones mediante la combinación de los distintos árboles en un modelo más sólido.

Agregación de bootstrap y criterios de selección

Los bosques aleatorios incorporan la aleatoriedad principalmente de dos maneras. Primero, seleccionando un subconjunto aleatorio de datos para cada árbol y, segundo, eligiendo un subconjunto aleatorio de variables predictoras para cada división del árbol. Cada uno de los árboles de un bosque aleatorio se genera utilizando una técnica conocida como agregación de *bootstrap*, o *bagging*. El algoritmo de *bagging* comienza por tomar múltiples muestras aleatorias del subconjunto original. Supongamos que tomamos B muestras, siendo B un número grande, generalmente en los cientos. Para cada muestra, se genera un extenso árbol de decisión, sin realizar ninguna simplificación. A continuación, estos árboles se usan para realizar predicciones. El algoritmo crea B reglas de predicción a partir de estos árboles y las combina. Para evaluar la precisión del modelo, se realizan B predicciones para cada punto de datos en base a los resultados de cada uno de los B árboles. El último paso es calcular un promedio de estas B predicciones para obtener el valor predicho final.

Los bosques aleatorios también fomentan la aleatoriedad limitando las características consideradas en cada división. En lugar de evaluar todas las variables predictoras (x variables) en cada ramificación, el algoritmo selecciona aleatoriamente un subconjunto de estas variables. Generalmente, el tamaño de este subconjunto es predeterminado y corresponde a la raíz cuadrada del número total de predictores, con un mínimo habitual de 4. Este enfoque se aplica a cada muestra de *bootstrap*, lo que conduce a la generación de B árboles. La predicción final se obtiene mediante el promedio de los resultados obtenidos de estos B árboles.

La razón de restringir el número de variables predictoras en cada división es minimizar la probabilidad de que todos los árboles resulten demasiado parecidos, especialmente si existe un

predictor predominante. Al limitar el conjunto de variables en cada punto de decisión, el algoritmo asegura una contribución más equilibrada de todos los predictores, incluyendo los más débiles, que pueden aportar información valiosa al considerarse de forma conjunta. Sin esta selección aleatoria, los árboles resultantes tenderían a privilegiar desproporcionadamente los predictores más fuertes, lo que conduciría a predicciones altamente correlacionadas y menos diversas.

Ajuste del modelo

Al ejecutar un bosque aleatorio, es necesario considerar varios parámetros de ajuste esenciales para asegurar un desempeño óptimo del modelo. Los principales parámetros son el número de árboles, el número de predictores evaluados en cada división y la regla de parada del crecimiento del árbol.

- **Número de árboles (B):**
  - Este parámetro determina cuántas muestras de *bootstrap* se utilizan para construir el bosque. Generalmente, entre más árboles, mayor será la precisión del modelo, pero también el tiempo de cálculo.
- **Número de predictores por división (x):**
  - En cada nodo, se selecciona solamente un subconjunto de predictores para la división. Una buena norma es utilizar la raíz cuadrada el número total de predictores. Por ejemplo, con 64 predictores, se deberían usar alrededor de ocho para cada división. No se deben usar menos de cuatro predictores.
- **Regla de parada del crecimiento del árbol:**
  - Determina cuándo se debe detener la división de nodos en un árbol. Lo ideal es determinar un número mínimo de observaciones por nodo terminal. Por lo general se utilizan entre cinco y 20 observaciones.

A continuación, el método examina la combinación de estos tres parámetros de ajuste que produce el menor error de predicción. Este se calcula mediante la **RMSE** (raíz del error cuadrático medio, por sus siglas en inglés), que nos indica el grado de desviación de nuestras predicciones respecto a los valores reales.

Otra métrica importante es el error *out-of-bag*, conocido por sus siglas en inglés **OOB**. Este indicador estima el desempeño del modelo. Al generar cada árbol en el bosque, el algoritmo selecciona aleatoriamente alrededor del 62,2 por ciento de los datos, dejando el 36,8 por ciento restante sin usar o «fuera de la bolsa». Estos datos *out-of-bag* que no se utilizan en la construcción de un árbol pueden servir para estimar la eficacia de ese árbol al evaluar su capacidad de predicción de los datos OOB. Calculando el promedio de estos errores OOB en todos los árboles del bosque, se obtiene una estimación confiable del desempeño del modelo, denominada tasa de error OOB. Esta técnica asegura que todos los puntos de datos se consideren en la evaluación del desempeño del modelo, brindando así un indicador sólido de su precisión sin necesidad de un conjunto de prueba adicional (Hartshorn 2016).

Importancia de las variables

En el método del bosque aleatorio, es fundamental comprender la importancia de cada variable predictora para interpretar el modelo y mejorar la exactitud de sus predicciones. Para ello, se recurre a una estrategia directa conocida como importancia de la permutación, que evalúa la importancia de las variables observando los cambios en la precisión de las predicciones al mezclar

aleatoriamente los valores de cada predictor. Luego, se compara el desempeño predictivo del modelo utilizando tanto los valores originales como los permutados de la variable, utilizando específicamente datos OOB. La importancia de la permutación se calcula midiendo el aumento en el error de predicción —ya sea el error cuadrático medio (MSE, por sus siglas en inglés) para las tareas de regresión o la tasa de error para las tareas de clasificación— cuando se permutan los valores de una variable en los datos OOB. Un aumento considerable en el error indica la importancia de la variable. Esta técnica no solamente identifica los predictores clave, sino que también detecta las interacciones complejas entre las variables. Dado que el algoritmo de bosque aleatorio selecciona subconjuntos aleatorios para cada división, es capaz de reconocer todos los predictores correlacionados como importantes si alguno de ellos contribuye de forma considerable al resultado (Cutler y otros 2012).

A3 Ejemplo

En esta sección, presentaremos un ejemplo para ilustrar las características del bosque aleatorio. Nos enfocaremos en desarrollar el modelo y la predicción, explicando cada paso, pero sin ofrecer ejemplos concretos.

Supongamos que tenemos una población de 100 contribuyentes. Cada contribuyente rellena una declaración de impuestos que incluye la base imponible (el monto sobre el cual se aplican los impuestos) e información complementaria. Supongamos que la información complementaria incluye diez variables. Podrían ser, entre otros, el monto destinado a los salarios de los trabajadores y a los costos de producción. Es importante resaltar que esta información complementaria no forma parte de la base imponible, pero puede ser útil para comprender cómo se llega al nivel de la base imponible.

De los 100 contribuyentes, 50 son seleccionados para una auditoría. Esto significa que, para esos 50 contribuyentes, también tenemos información sobre (posibles) discrepancias entre lo declarado como base imponible y el monto real. A modo de aclaración, consideremos que cada uno de los 50 contribuyentes evade impuestos: a través de las auditorías, obtendremos (como mínimo) el monto declarado erróneamente y la base imponible real.

El primer paso consiste en comprender que solamente disponemos de información sobre la declaración de datos inexactos de 50 contribuyentes. Esto significa que solo podremos comparar las predicciones con las variables reales en este subconjunto para evaluar la exactitud del modelo de predicción. Por esta razón, segmentaremos la muestra total y nos enfocaremos en los contribuyentes que fueron auditados.

Dividimos la muestra de los contribuyentes auditados en dos submuestras. Para simplificar, seleccionaremos 25 contribuyentes para la muestra de entrenamiento y el resto para la muestra de prueba. Ejecutaremos un modelo de bosque aleatorio en la muestra de entrenamiento y luego utilizaremos la muestra de prueba para ajustarlo. Debemos determinar dos valores clave: el número de iteraciones (o árboles) y el número de predictores por división. La finalidad del modelo es estimar los montos declarados de forma inexacta a partir de las covariables (las diez variables adicionales en las declaraciones de las empresas). Por ende, utilizaremos todas las variables por dos razones principales. Primero, porque esas variables nos ayudan a caracterizar la base imponible, lo que resulta relevante para determinar su nivel. Segundo, porque ya que contamos con esta información, también son determinantes para decidir qué contribuyente será sometido a auditoría. Incluir todas las variables nos permite evitar un sesgo de selección muestral por factores observables.

Comencemos por establecer cuántos árboles necesitamos. Para ello, conservamos el número de predictores utilizados en cada división (variables utilizadas para la estimación de la declaración de datos inexactos). Para simplificar, supongamos que vamos a usar una de las diez variables disponibles. Para decidir el número de árboles, debemos ejecutar el modelo en la muestra de entrenamiento y comparar las predicciones en la muestra de prueba utilizando distintas cantidades de árboles. En otras palabras, ejecutamos N veces N bosques aleatorios distintos, cambiando únicamente el número de árboles que utilizamos. Luego, ejecutamos el modelo para realizar predicciones en la muestra de prueba y comparamos los valores predichos con las irregularidades reales detectadas durante el proceso de auditoría. Al final, tendremos N valores de la RMSE (una por cada ejecución del modelo). Seleccionamos el valor más bajo y comprobamos el número de árboles correspondiente, B. Este es el número óptimo de árboles porque reduce al mínimo el error de predicción indicado por la RMSE, es decir que produce la estimación más acertada de la base imponible declarada de forma inexacta.

Ahora, procedemos a estimar el predictor utilizado en cada división. En este caso, utilizamos el número óptimo de árboles, B, que determinamos en el proceso anterior. Repetimos el mismo proceso iterativo, pero esta vez, ejecutamos diez modelos de bosques aleatorios distintos, obtenemos la predicción de cada uno de ellos en la muestra de entrenamiento y la comparamos con el valor real declarado de forma inexacta. Ejecutamos diez modelos porque tenemos diez variables disponibles. La razón es que el número total de variables es el número máximo de predictores para cada división. Es importante destacar que, si se cuenta con diez variables, pero se opta por utilizar solamente ocho para el modelo de predicción, se deberán ejecutar ocho modelos. La cantidad de modelos a ejecutar en este paso debe ser siempre igual a la cantidad de variables seleccionadas para el modelo predictivo. Por último, repetimos el proceso seleccionando la RMSE más baja y comprobando el número de predictores utilizados, *x*. Este número de predictores *x* es el óptimo para reducir el error de predicción al mínimo.

A través de estos dos pasos, encontramos el número ideal de árboles (B) y de predictores por división (x) para utilizar en el bosque aleatorio. Recordemos que, para calcularlos, utilizamos los 50 contribuyentes auditados, dividiendo la muestra en un conjunto de entrenamiento y un conjunto de prueba. Ahora, podemos realizar predicciones para los otros 50 contribuyentes que no fueron sometidos a auditoría. El procedimiento es el siguiente. Primero, ejecutamos el bosque aleatorio con los parámetros óptimos en el conjunto de los 50 contribuyentes auditados. Luego, predecimos los valores en el conjunto de los 50 contribuyentes no auditados. Por último, podemos crear una variable que incluya tanto la declaración de datos inexactos descubierta en los 50 contribuyentes auditados como la predicción de declaración de datos inexactos para los 50 contribuyentes no auditados.