Estudios de la OCDE sobre Gobernanza Pública

La Lucha contra el Fraude en las Subvenciones Públicas en España

APRENDIZAJE AUTOMÁTICO PARA EVALUAR LOS RIESGOS Y ORIENTAR LAS ACTIVIDADES DE CONTROL

OCDE

Estudios de la OCDE sobre Gobernanza Pública

La Lucha contra el Fraude en las Subvenciones Públicas en España

APRENDIZAJE AUTOMÁTICO PARA EVALUAR LOS RIESGOS Y ORIENTAR LAS ACTIVIDADES DE CONTROL

OCDE
MEJORES POLÍTICAS PARA UNA VIDA MEJOR

El proyecto fue cofinanciado por la Unión Europea a través del Programa de Apoyo a las Reformas Estructurales (REFORM/IM2020/006). Esta publicación se hizo con la ayuda financiera de la Unión Europea. Las opiniones expresadas en el presente documento no pueden en modo alguno tomarse como un reflejo de la opinión oficial de la Unión Europea.

Por favor, cite esta publicación de la siguiente manera:
OECD (2021), *La Lucha contra el Fraude en las Subvenciones Públicas en España: Aprendizaje Automático para Evaluar los Riesgos y Orientar las Actividades de Control*, Estudios de la OCDE sobre Gobernanza Pública, OECD Publishing, Paris, *https://doi.org/10.1787/6a4ab581-es*.

ISBN 978-92-64-52745-4 (impresa)
ISBN 978-92-64-49803-7 (pdf)

Estudios de la OCDE sobre Gobernanza Pública
ISSN 2414-3308 (impresa)
ISSN 2414-3316 (en línea)

Imágenes: Portada © 2-Q STOCK/Shutterstock y Gearings imagen © OCDE, diseñada por Christophe Brilhault.

Las erratas de las publicaciones se encuentran en línea en: *www.oecd.org/about/publishing/corrigenda.htm*.

Prólogo

El fraude en los programas de subvenciones públicas desvía el dinero de los contribuyentes apartándholo de los servicios esenciales y reduce los beneficios para los beneficiarios bien intencionados. Cuando los beneficiarios individuales, los proveedores privados o los funcionarios públicos cometen fraudes en los programas de subvenciones, no solo socavan la integridad del programa en sí, sino que también se corre el riesgo de erosionar la confianza en las administraciones públicas. A raíz de la pandemia de COVID-19, marcada por un alto volumen de gasto acelerado, los riesgos de fraude se han convertido en una preocupación apremiante para los gobiernos de todo el mundo.

En este entorno, los organismos públicos de control y auditoría desempeñan un papel fundamental para garantizar que el dinero se gasta bien y que las vulnerabilidades se detectan y abordan rápidamente. En España, la Intervención General de la Administración del Estado (IGAE) está a la vanguardia de los esfuerzos para prevenir y detectar el fraude. Su mandato de supervisión se centra en áreas de alto riesgo en las que el fraude suele emboscarse, como en el caso de este informe, en la concesión de subvenciones públicas.

Adoptar un enfoque basado en el riesgo es fundamental para canalizar unos recursos limitados. Con este fin, los organismos de control modernos, como la IGAE, dependen cada vez más de los datos y la analítica como herramientas fundamentales para evaluar los riesgos. Al aprovechar los datos y el análisis para mejorar los procesos de control y auditoría, la IGAE estará mejor dotada para identificar los riesgos y dirigir sus recursos donde tengan mayor impacto. Este informe refleja la iniciativa y el compromiso de la IGAE para aprovechar enfoques de vanguardia, incluidas metodologías punteras en inteligencia artificial y aprendizaje automático.

El presente documento fue revisado por el Grupo de Trabajo de Altos Funcionarios de Integridad Pública (SPIO) de la OCDE en 1 de noviembre de 2021 y desclasificado por el Comité de Gobernanza Pública en 23 de noviembre de 2021. Fue elaborado para que lo publicara la Secretaría de la OCDE. El proyecto fue cofinanciado por la Unión Europea a través del Programa de Apoyo a las Reformas Estructurales (REFORM/IM2020/006). Esta publicación se hizo con la ayuda financiera de la Unión Europea. Las opiniones expresadas en el presente documento no pueden en modo alguno tomarse como un reflejo de la opinión oficial de la Unión Europea.

Agradecimientos

Bajo la dirección de Elsa Pilichowski, directora de Gobernanza Pública de la OCDE, y Julio Bacio Terracino, director interino de la División de Integridad del Sector Público, este proyecto fue dirigido por Gavin Ugale, quién redactó y editó el Capítulo 1. El Dr. Mihaly Fazekas, profesor asistente de la Central European University y director científico del Instituto de Transparencia del Gobierno, diseñó el modelo de aprendizaje automático descrito en el Capítulo 2 y redactó el Capítulo 3, con el apoyo de Viktoriia Poltoratskaia. Varun Banthia apoyó la investigación y el proceso de redacción. Meral Gedik, Andrea Uhrhammer, Laura Völker y Elisabeth de Vega Alavedra proporcionaron asistencia editorial. Charles Victor y Aman Johal proporcionaron asistencia administrativa y Balazs Gyimesi proporcionó apoyo para las comunicaciones.

La OCDE agradece a los colegas de la Intervención General de la Administración del Estado español (IGAE) su fructífera cooperación y liderazgo. En concreto, la OCDE quiere dar las gracias a Isabel Silva Urien e Ismael García Cebada, así como a sus equipos, incluidos Carlos Collado Molinero, Pablo Lanza Suárez e Israel Barroso Pérez. La OCDE también quiere dar la gracias a Ciresica Feyer, de la Dirección General de Apoyo a las Reformas Estructurales de la Comisión Europea (DG REFORM) por su orientación a lo largo del proyecto y sus aportes al borrador del informe.

Índice

FIGURAS

TABLAS

Abreviaciones y acrónimos

AEAT	Agencia Estatal de Administración Tributaria
BDNS	Base de Datos Nacional de Subvenciones
COSO	Comité de la Organización Patrocinadora de la Comisión Treadway
UE	Unión Europea
IGAE	Intervención General de la Administración del Estado
OIP	Oficina de Informática Presupuestaria
OLAF	Oficina Europea de Lucha contra el Fraude
ONA	Oficina Nacional de Auditoría
ONC	Oficina Nacional de Contabilidad
SAIs	Auditorías Generales Nacionales
SHAP	Explicaciones de aditivos SHapley
SNPSAP	Sistema Nacional de Publicidad de Subvenciones y Ayudas Públicas
TGSS	Tesorería General de la Seguridad Social

Resumen ejecutivo

El fraude es por naturaleza una actividad oculta. Entonces, ¿cómo pueden las autoridades detectar y mitigar los riesgos de manera efectiva? Este informe identifica vías para que la Intervención General de la Administración del Estado (IGAE) para hacer frente a este desafío, utilizando modelos de aprendizaje automático de última generación y orientar eficazmente sus actividades de control a los mayores riesgos de fraude que se encuentran en subvenciones y subsidios públicos.

Hay pocas cifras fiables del nivel de fraude a escala nacional, dadas las complejidades de medir algo que está oculto intencionadamente. A menudo, los países se basan en mediciones indirectas más amplias, como el alcance de las irregularidades notificadas en programas o sectores específicos. No obstante, las cifras disponibles sugieren desafíos importantes y riesgos de fraude para las administraciones públicas. Por ejemplo, en países que evalúan el alcance del fraude en los programas de beneficios sociales, como Francia, el Reino Unido y Estados Unidos, las estimaciones de fraude alcanzan los cientos de millones de euros. En su 32º Informe Anual sobre la protección de los intereses financieros de la Unión Europea: Lucha contra el fraude 2020, la Comisión Europea informó de 375 millones de euros fraudulentos vinculados a ingresos y gastos. Es probable que los niveles de fraude en los Estados miembros de la UE sean mucho más altos si se tienen en cuenta además los fondos nacionales y el gasto público.

Los organismos de control, como la IGAE, están en primera línea de los esfuerzos de la administración pública para prevenir y detectar el fraude. Tienen una visión exclusiva en toda la administración para detectar los riesgos de fraude y fortalecer la eficacia, eficiencia y economía del gasto público a través de evaluaciones previas y posteriores. Para hacer este trabajo de manera eficaz en la era digital, los órganos de supervisión se enfrentan a una presión considerable para no perder el tren de la evolución de los riesgos y de las nuevas tecnologías. En España, al igual que otros Estados miembros de la UE, el Plan de Recuperación, Transformación y Resiliencia hace especial hincapié en la necesidad de mejorar los mecanismos y herramientas para prevenir, detectar y corregir los riesgos inherentes en las subvenciones públicas, incluidos el fraude, la corrupción, los conflictos de interés y la doble financiación.

En este contexto, la IGAE y la OCDE, con el apoyo de la Comisión Europea, han trabajado juntas para identificar los métodos con los que la IGAE pueda fortalecer sus evaluaciones de riesgos de fraude en ayudas y subvenciones públicas, con el objetivo final de implantar actividades de control mejor focalizadas. El proyecto se ha centrado en asesorar a la IGAE en hacer uso de los datos existentes e identificar vías para ampliar su análisis y tener en cuenta nuevas fuentes de datos, riesgos de fraude y metodologías. El Capítulo 1 describe brevemente el contexto y el mandato de la IGAE, así como su enfoque en la evaluación de riesgos y planificación de sus actividades de control. También se efectúan varias consideraciones generales para que la IGAE mejore su uso de datos y análisis, independientemente de si adopta el modelo de aprendizaje automático referido en el Capítulo 2, o no, con un enfoque en la evaluación de riesgos de fraude de subvenciones. Esto incluye:

- Fortalecer la gobernanza y la gestión de datos para evaluar los riesgos de fraude de subvenciones, comenzando con mejoras de fácil implementación, como mejorar los diccionarios de datos, la claridad de los identificadores únicos y los controles de datos específicamente para el análisis de riesgos de fraude.

- Desarrollar la capacidad de evaluar riesgos basándose en datos, en concreto, desarrollando conjuntos de datos estructurados e idealmente una capacidad que reúna la experiencia relacionada con procesos de concesión de subvenciones, riesgos de fraude, análisis y visualización.
- Ser consciente de las trampas relacionadas con los indicadores compuestos de riesgo, así como con los posibles sesgos, que pueden incluir sesgos en los modelos de aprendizaje automático.

El Capítulo 2 presenta una prueba de concepto para un modelo de riesgo basado en datos, para que la IGAE lo adopte en parte o en su totalidad. La metodología utiliza datos actualmente a disposición de la IGAE, por lo que implícitamente reconoce el contexto presente de la IGAE. El modelo de aprendizaje automático tiene en cuenta los riesgos a lo largo del ciclo de la subvención en la medida en que lo permiten los datos. El proceso de desarrollo de la prueba de concepto para el modelo de riesgo ha desvelado distintos conocimientos y la identificación de áreas de mejora, entre ellas:

- Establecer un conjunto de datos listo para identificar riesgos de fraude, que este proyecto ha comenzado como piloto y puede sentar bases para el análisis de riesgos futuros con menor inversión en recursos y tiempo.
- Ampliar el uso de indicadores por parte de la IGAE a lo largo de todo el ciclo de subvenciones, incluida la mejora de datos e indicadores que van más allá de las características descriptivas y revelan comportamientos (por ejemplo, conflictos de interés).
- Invertir en la mejora continua del modelo de riesgo de aprendizaje automático, si se adopta, para garantizar una muestra verdaderamente aleatoria, teniendo en cuenta los nuevos datos y riesgos y abordar los sesgos, entre otras consideraciones.
- Tener en cuenta los análisis de red y hacer uso de un conjunto más amplio de metodologías, incluidas aquellas que aprovechan los datos mercantiles de las empresas.

Por último, el Capítulo 3 ofrece una hoja de ruta para completar los datos existentes sobre subvenciones de la IGAE a fin de mejorar sus modelos de evaluación de riesgos. Concretamente, describe conjuntos de datos que pueden integrarse con los datos existentes de subvenciones en la IGAE, mejorando así la sofisticación analítica y la precisión de la evaluación de riesgos. La orientación y las recomendaciones del informe se basan en entrevistas de investigación de la OCDE, análisis del contexto de la IGAE y los datos disponibles, las experiencias de otras entidades públicas y las principales prácticas internacionales.

1 Control basado en riesgos en España: una base para fortalecer los análisis

Este capítulo ofrece una visión general de la Intervención General de la Administración del Estado (IGAE) y su supervisión de las subvenciones y subsidios públicos en España. Describe el enfoque actual de la IGAE para la planificación basada en el riesgo y destaca las condiciones previas y las consideraciones para que la IGAE avance en el uso de datos de subvenciones para evaluar los riesgos de fraude. Esto incluye consideraciones y recomendaciones para garantizar una gobernanza y una gestión de datos efectivas, así como desarrollar la capacidad para utilizar modelos de aprendizaje automático.

Introducción

La Intervención General de la Administración del Estado (IGAE) lleva a cabo el control interno sobre la gestión económica y financiera de la administración del estado español. Esto incluye la administración general del Estado, los organismos autónomos dependientes, las entidades estatales de derecho público y las entidades públicas empresariales. Como parte de su mandato, la IGAE ejecuta actividades de control para garantizar una buena gestión financiera y el cumplimiento, entre otras, de la Ley Orgánica de Estabilidad Presupuestaria y Sostenibilidad Financiera, la Ley General de Subvenciones y la legislación de la Unión Europea (OCDE, 2014[1]). La IGAE también investiga áreas de alto riesgo de posibles fraudes e irregularidades, incluidas las subvenciones públicas y las ayudas que apoyan la consecución de los objetivos de las políticas públicas de España.[1]

Las ayudas y subvenciones públicas que supervisa la IGAE ascienden a 89 860 millones de euros del presupuesto total anual, e involucran a miles de beneficiarios y entidades. Dado el tamaño de este universo auditor y el gran volumen de operaciones relacionadas con el pago de ayudas, la IGAE ha desarrollado un enfoque basado en el riesgo como herramienta para abordar los riesgos más elevados y administrar sus recursos de manera eficiente. Los criterios de riesgo que la IGAE ha desarrollado tienen en cuenta el potencial de fraude e irregularidad en función de criterios predeterminados, como se describe en este capítulo.

Si bien la IGAE ha desarrollado un enfoque basado en el riesgo para sus actividades de control, existen oportunidades de hacer un mejor uso de los datos existentes y de las nuevas metodologías para centrar aún mejor sus recursos en áreas de alto riesgo. Este capítulo explora las consideraciones claves para que la IGAE avance en el uso de datos y análisis. Como se describe en el Capítulo 2, el proyecto se ha centrado en una metodología específica, inspirada en el aprendizaje automático. En cualquier caso, las consideraciones de este capítulo son de aplicación general, independientemente de la técnica o metodología a emplear. Además, si bien el proyecto de la OCDE ha puesto el foco en mejorar la detección de riesgos de fraude de subvenciones, los conocimientos de este capítulo y el siguiente son aplicables a otros tipos de análisis de riesgo cuando se disponga de datos fiables.

Descripción general del ciclo de subvenciones y las responsabilidades de supervisión de la IGAE

La IGAE sigue un modelo de funcionamiento descentralizado, con tres funciones de servicio central que desarrollan sus áreas de responsabilidad a nivel de la administración central, incluyendo la Oficina Nacional de Auditoría (ONA), la Oficina Nacional de Contabilidad (ONC) y la Oficina de Informática Presupuestaria (OIP). La IGAE tiene responsabilidades de control previo y posterior:

- Previo, controlando, antes de su aprobación, las tareas de ejecución de gastos, ingresos, pagos e inversiones, o la aplicación general de fondos públicos, para garantizar que la administración cumpla todas las leyes vigentes. El control previo tiene por tanto un carácter preventivo, que se realiza antes de la concreción de diversos actos económicos, como contratos, ayudas, convenios, pagos y nóminas, entre otros. Puede ejercerse de forma limitada, examinando determinados aspectos críticos de los actos económicos y financieros, o puede abarcar a su totalidad, examinando toda la documentación vinculada a un acto.

Posterior, mediante la verificación permanente del estado y funcionamiento de las entidades del sector público para comprobar el cumplimiento de la normativa aplicable y que la gestión se ajusta a los principios de buena gestión financiera, y en particular, a la consecución del objetivo de estabilidad presupuestaria y financiera. La IGAE lleva a cabo auditorías públicas, que pueden adoptar diversas formas, incluyendo auditorías anuales de regularidad contable (revisión de la información contable para verificar su

cumplimiento de las normas contables), auditorías de cumplimiento normativo (verificación de la legalidad de la gestión presupuestaria, adquisiciones, personal, gestión de ingresos y ayudas) y auditorías de desempeño (examinar operaciones y procedimientos para evaluar la racionalidad financiera y económica y el cumplimiento de los principios de buena gobernanza como medio para detectar deficiencias y hacer recomendaciones para subsanarlas).

Los principales resultados de las auditorías de la IGAE se resumen en un informe anual. Cuando se detectan infracciones que pudieran derivar en corrupción o fraude, se envía un informe especial al Ministerio de Hacienda y Función Pública, además de a la entidad controlada. Este informe propicia mejoras a lo largo del tiempo en las técnicas y procedimientos de gestión económica y financiera, a medida que se actúa siguiendo sus recomendaciones. Existen mecanismos de colaboración entre la IGAE, intervenciones de las comunidades autónomas e intervenciones locales (OCDE, 2014[1]).

El mandato de control de las ayudas recae principalmente sobre la ONA y la División de Control e Información de Subvenciones de la IGAE. Sin embargo, hay Intervenciones Delegadas a nivel regional o provincial e Intervenciones Delegadas integradas en los ministerios y en las organizaciones del sector público. Estas entidades actúan como controladores financieros y son responsables del seguimiento continuo de los controles financieros y las auditorías internas públicas (IGAE, 2020[2]). Además, las Intervenciones Delegadas tienen la labor de controlar los gastos públicos dirigidos a terceros, incluyendo las subvenciones públicas, préstamos y garantías.

El artículo 140.2 de la Ley 47/2003, General Presupuestaria, otorga a la IGAE la facultad de ejecutar el control interno del sector público con plena autonomía frente a las autoridades y demás entidades cuya gestión controla (Gobierno de España, 2003[3]). Esta facultad incluye la autoridad de ejecutar actividades de control relacionadas con los beneficiarios de las ayudas, de acuerdo con los artículos 141 y 140.2 de la Ley 47/2003 y (Gobierno de España, 2003[3]) y 44 de la Ley 38/2003, (Gobierno de España, 2003[4]) General de Subvenciones (IGAE, 2020[5]).

La administración, o el organismo que concede la subvención, supervisa los procesos generales en cada fase del ciclo de la subvención, y el organismo que concede la subvención es responsable de supervisar al beneficiario para garantizar el cumplimiento de las condiciones de la subvención. Por ejemplo, al inicio del ciclo de la subvención, los asuntos que se abordan pueden incluir si la agencia que concede la subvención ha generado correctamente la subvención y si la subvención se concedió, aplicó y verificó con exactitud. Además de la supervisión de la agencia concedente, los organismos externos, como las Cortes, el Tribunal de Cuentas y otros órganos de auditoría, proporcionan supervisión y controles adicionales. El modelo de la IGAE para la supervisión de las subvenciones públicas abarca el ciclo de subvenciones que, normalmente, consta de las siguientes fases:

1. *Competencia*: las condiciones que debe cumplir el beneficiario de una subvención para recibir esta son definidas por la agencia concedente. La agencia que otorga la subvención aprueba estas condiciones y se abre un periodo de solicitud.
2. *Selección*: los candidatos se revisan y seleccionan en función de la calidad de sus solicitudes en comparación con el conjunto de criterios original.
3. *Ejecución de la subvención*: si el solicitante ya cumple todos los requisitos para la subvención, o si la subvención tiene una provisión para un anticipo, se realiza un pago y el beneficiario debe comenzar inmediatamente las actividades exigidas por la subvención.
4. *Seguimiento*: una vez que se cumplen los requisitos de la subvención, el beneficiario debe presentar una justificación de cómo se han gastado los fondos. La agencia concedente revisará esta justificación y decidirá si es necesario realizar algún pago final o si es necesario recuperar los fondos. Esto último puede tener lugar si una actividad no se completó según lo estipulado en el acuerdo de concesión inicial.

Las actividades de investigación y control de la IGAE a lo largo del ciclo de subvenciones tienen diferentes propósitos. Por ejemplo, su objetivo es verificar si el beneficiario obtuvo y ha gestionado correctamente la subvención. La IGAE también puede evaluar si la subvención estaba justificada y si las operaciones cubiertas por la subvención son reales y legítimas. La IGAE también investigará si el beneficiario no informó a la administración de hechos materiales que pudieran haber afectado a la concesión de la subvención. (IGAE, 2020[5])

La transparencia se enfatiza en las leyes y en la práctica. Por ejemplo, el Real Decreto 130/2019 (Gobierno de España, 2019[6]) integra muchas de las normas mencionadas y reitera las disposiciones relativas a la transparencia, el acceso a la información pública y la buena gobernanza. Esta ley, junto con el Reglamento (UE) n.º 651/2014 (Unión Europea, 2014[7]) y el 702/2014, (Unión Europea, 2014[8]) dicta que los datos sobre estas subvenciones y su desembolso deben ser publicados cada año, sin restricciones, en el Sistema Nacional de Publicidad de Subvenciones y Ayudas Públicas (SNPSAP) (IGAE Ministerio de Hacienda,, 2021[9]).

El enfoque de la IGAE para la planificación basada en riesgos

Anualmente la IGAE elabora un plan de controles previos y posteriores. Este plan se dirige a qué controles abordan riesgos más elevados y cuáles contribuyen de manera más eficaz al avance de los cuatro objetivos generales del organismo. Estos objetivos son: 1) Combatir el fraude; 2) Aumentar la sensibilización de las actividades de control entre los beneficiarios y organismos que conceden las subvenciones; 3) Buscar un valor añadido en el control, más allá de la simple verificación o repetición; y 4) Cumplir los principios de descentralización utilizando todos los recursos, medios y herramientas disponibles para las actividades de control. Los controles planificados previamente, que estaban incompletos del año anterior, también se transferirán al plan anual. Los planes anuales de la IGAE están sujetos a cambios a lo largo del año si surgen nuevos riesgos imprevistos (IGAE, 2020[5]). Por ejemplo, en 2021, la IGAE seleccionó dos controles para evaluar adicionalmente: uno de los cuales es que no se hubieran concedido subvenciones a ninguna entidad inhabilitada y otro es que no se hubieran concedido subvenciones que superen los umbrales regulatorios de la Comisión Europea (IGAE, 2020[5]).

La IGAE suele planificar sus actividades de control en función del análisis de la Base de Datos Nacional de Subvenciones (BDNS), que es una base de datos que tiene información sobre las subvenciones de todas las administraciones públicas y sus beneficiarios, y está administrada por la IGAE. La IGAE también se apoya en CincoNet y Presya, que son respectivamente el sistema contable de la administración del estado y el sistema de contabilidad de préstamos públicos, así como en denuncias. Ejemplos de fuentes de denuncia son la Agencia Estatal de Administración Tributaria (AEAT), delatores individuales, organizaciones concedentes e investigadores de blanqueo de capitales. La información sobre el beneficiario real está disponible en una variedad de fuentes (por ejemplo, la base de datos del Consejo General del Notariado), y el Ministerio de Justicia está desarrollando un Registro de Titularidad efectiva que consolidará diversas fuentes. La experiencia de años anteriores y el conocimiento contextual también ayudan a la IGAE a determinar qué áreas son de alto riesgo y tienen debilidades de control.

Los objetivos, prioridades y limitaciones de recursos de la IGAE también se tienen en cuenta para planificar las actividades para el próximo año (IGAE, 2020[5]). Para promover el uso eficiente de sus recursos, la IGAE utiliza una aproximación basada en el riesgo que describe en su Plan de Auditorías y Control Financiero de Subvenciones y Ayudas Públicas para 2021, incorporando los marcos de control reconocidos internacionalmente, como el del Comité de la Organizaciones Patrocinadoras de la Comisión Treadway (COSO) y la Estrategia Antifraude de la Comisión Europea. El plan subraya las tres consideraciones principales de la IGAE:

1. Subvenciones con mayor riesgo percibido: como indicadores de riesgo, la IGAE tiene en cuenta el importe concedido, el nivel de fraude observado en años anteriores, las características de las convocatorias de subvenciones y de los procedimientos de concesión, justificación y verificación.
2. La visibilidad del control: la IGAE tiene en cuenta la visibilidad y el impacto de la actividad de control, reconociendo que las actividades de gran visibilidad pueden actuar como elemento disuasorio (es decir, los beneficiarios y otras partes interesadas son más conscientes de la vigilancia de la IGAE) y pueden propiciar una mejor gestión.
3. La «rentabilidad de los medios disponibles»: se refiere en términos generales a la consideración de la IGAE de la eficiencia de sus actividades de control y la estructura descentralizada (Servicios Periféricos), que incluye la colaboración con los ministerios y departamentos regionales (IGAE, 2020[5]).

En la planificación y ejecución de su trabajo, la IGAE debe seguir determinados parámetros que configuran sus actividades de control. Su mandato se limita al control de las subvenciones y ayudas públicas, incluyendo los préstamos, contemplados en el Título III de la Ley General de Subvenciones (LGS). Además, las actividades de control de la IGAE para 2021 se centran en el ejercicio económico de 2018 o posterior, reconociendo que algunas subvenciones tienen períodos de ejecución de varios años. Las actividades de control de la IGAE se limitan principalmente a subvenciones y ayudas financiadas con fondos nacionales, aunque es posible que una acción subvencionada también haya recibido financiación de la Unión Europea (UE) (IGAE, 2020[5]).

Los funcionarios de IGAE destacan tres áreas claves de riesgo de fraude que preocupan especialmente en los programas públicos de subvenciones en España: 1) Facturación excesiva de horas por parte de los beneficiarios; 2) Doble financiación; y 3) Exceso de facturación por parte de contratistas o terceros.

- Existe el riesgo de que los beneficiarios facturen horas extra sobre el servicio real prestado. Las organizaciones que reciben fondos de subvenciones deben informar a la agencia concedente de cuántas horas de trabajo realiza el personal en el proyecto subvencionado. Esta cifra tiene implicaciones sobre la cantidad de fondos que recibe el beneficiario. Sin embargo, dado que las operaciones de muchas organizaciones se financian solo en parte mediante subvenciones, existe el riesgo de que se inflen las horas de trabajo realizadas por los empleados en trabajos relacionados con las subvenciones. Por ejemplo, una organización podría afirmar falsamente que los costes salariales en los que se habría incurrido, aún en ausencia de cualquier subvención eran, en cambio, un resultado directo de la financiación pública (ver en el Recuadro 1.1 a continuación, la experiencia de los Centros de Servicios de Medicare y Medicaid de EE. UU.). La IGAE intenta controlar este riesgo exigiendo informes laborales sobre las horas empleadas y aplicando umbrales máximos; sin embargo, estos enfoques solo pueden mitigar parcialmente los riesgos. Los funcionarios de la IGAE destacan la necesidad de mejorar los datos para ayudar a detectar este tipo de fraude, incluyendo los datos de salario/hora, los ingresos totales de la empresa y los gastos de personal habituales previos a la concesión de la subvención. Esta información podría añadirse al BDNS para respaldar un análisis adicional, según los funcionarios. Esto podría incluir la comparación de los beneficiarios con sus competidores u operadores del mismo sector para encontrar aquellos que usan de manera ineficiente las horas de trabajo, o comprobaciones previas y posteriores para evaluar las discrepancias entre lo que una empresa afirma en los documentos de la subvención y los salarios que realmente paga.
- Una segunda área de preocupación es que los beneficiarios podrían recibir financiación de dos o más fuentes, tanto públicas como privadas, a un nivel que exceda los costes incurridos y se traduzca en ganancias indebidas. La BDNS ayuda a la IGAE a combatir esta práctica, ya que incluye una lista de todas las subvenciones de todas las administraciones concedidas a una sola organización. Sin embargo, el BDNS no incluye las subvenciones de la UE. Los funcionarios de la IGAE señalan que disponer de esta información adicional sobre todas las subvenciones otorgadas

a una organización y los ingresos totales de cada una de las fuentes sería especialmente útil para identificar áreas de alto riesgo. Hoy en día esta declaración omnicomprensiva de todas las fuentes de financiación ya es un requisito para los grandes beneficiarios, pero no lo es para los más pequeños. Se podría lograr una extensión de esta declaración obligatoria a todos los beneficiarios a través de una autodeclaración por parte del beneficiario, una búsqueda en la web o un análisis de los estados financieros de la organización.

- El exceso de facturación o subcontratación se produce cuando un proveedor del beneficiario cobra de más por un servicio o suministro en particular, ya sea cobrando un valor superior al de mercado o proporcionando la cantidad inferior a la indicada. Este riesgo de fraude es especialmente difícil de detectar, ya que a menudo va acompañado de un rastro documental legítimo. Los funcionarios de la IGAE destacan la necesidad de aprovechar las nuevas tecnologías y datos para identificar los casos en los que esto puede ocurrir y para analizar mejor el entorno en el que operan las empresas y sus proveedores, el contexto geográfico y las relaciones entre ellos. De hecho, analizar las relaciones puede ser útil para un análisis de riesgos más amplio, yendo más allá del análisis del exceso de facturación por sí solo. Esto puede incluir relaciones entre proveedores, beneficiarios, filiales de los beneficiarios o empresas relacionadas y organizaciones concedentes. Estos tipos de relaciones pueden dar lugar a una apropiación indebida o a que se conceda a las organizaciones un exceso de financiación. Los funcionarios de la IGAE señalan que crear una base de datos que rastree este tipo de relaciones sería útil para identificar áreas de alto riesgo de fraude. Véase el capítulo 2 para ver un ejemplo y una exposición más detallada de las técnicas de análisis de redes.

Recuadro 1.1. Atacando la sobrefacturación por parte de los Centros de Servicios de Medicare y Medicaid de EE. UU.

Cuando las administraciones públicas financian a terceros, un área de riesgo frecuente es que el beneficiario facture horas de trabajo en exceso, ya sea por error o con mala intención. En Estados Unidos, los Centros de Servicios de Medicare y Medicaid (CMS) utilizan un sistema de análisis predictivo para intentar detectar sobrefacturaciones de este tipo. El Sistema de Prevención de Fraude (FPS) utiliza diversos datos para evaluar una serie de métricas, entre ellas:

- Diferenciación basada en reglas, como la identificación de tarjetas de crédito o cuentas asociadas a comportamientos fraudulentos en el pasado.
- Identificación de anomalías, como señalar beneficiarios que comparativamente facturan importes mayores que otros similares.
- Análisis predictivo, que identifica a los beneficiarios que tienen características similares a conocidos agentes de conducta negativa.
- Análisis de redes, mediante el cual se comparan números de teléfono y direcciones de beneficiarios con los de conocidos agentes de conducta negativa.

Al examinar estos rasgos y comportamientos, los CMS han identificado una serie de entidades con prácticas de facturación de alto riesgo. Después de haber sido marcadas por el programa de análisis y después de aplicar otras formas más tradicionales de investigación, se han bloqueado varias entidades para que no sigan facturando o para inhabilitarlas en el sistema. El FPS ha permitido a los CMS asignar sus recursos de manera eficiente y eficaz. En última instancia, se estima que el programa ha ahorrado a los contribuyentes más de 200 millones de USD, que es una rentabilidad de 5 USD por cada 1 USD de inversión en el sistema.

Fuente: (Centers for Medicare & Medicaid Services, 2014[10])

Consideraciones comunes para el uso de datos y análisis para evaluar riesgos

La IGAE es fundamentalmente un consumidor de datos, ya que depende de los inputs de datos de otras entidades públicas para realizar su trabajo de supervisión y evaluar riesgos. Como se ha expuesto, gran parte de estos datos se integran en la BDNS, pero la IGAE también hace uso de otras fuentes, como sistemas contables, bases de datos de préstamos y datos de denuncias. La IGAE también mantiene sus propios registros sobre el resultado de las actividades de control y casos sancionados. Los funcionarios de la IGAE destacan la utilización de controles de calidad y controles destinados a garantizar la fiabilidad de los datos que utilizan. Sin embargo, a la vez que ha apoyado a la IGAE para desarrollar la metodología de riesgos detallada en el Capítulo 2, la OCDE ha identificado áreas en las que la IGAE podría tomar medidas adicionales para mejorar su uso de datos y análisis, independientemente de la técnica o metodología concreta. En términos generales, como se detalla en esta sección, esto incluye: 1) Mejoras en la gobernanza y gestión de datos de la IGAE; 2) desarrollar aún más su capacidad de análisis utilizando datos y análisis; y 3) tener en cuenta los errores relacionados con los métodos avanzadas de evaluación de riesgos, como las limitaciones de uso de indicadores de riesgo compuestos y sesgos.

Fortalecer la gobernanza y la gestión de datos

La gobernanza de datos, y más concretamente la gestión de datos, es la piedra angular de un análisis eficaz, incluido el enfoque descrito en el Capítulo 2. Independientemente de la metodología específica, cualquier enfoque «basado en datos» se fundamenta en estos elementos. El modelo descrito en Figura 1.1 destaca los valores de todos los aspectos organizativos, políticos y técnicos para una gobernanza de datos exitosa.

Figura 1.1. Gobernanza de datos en el sector público

Fuente: (OCDE, 2019[11])

El modelo de gobernanza de datos anterior es relevante tanto desde una perspectiva institucional como de toda la administración pública. Las instituciones de auditoría, respecto a la gobernanza y la gestión de datos están a la vanguardia en su trabajo diario. Las normas y directrices internacionales, en concreto las avanzadas por las instituciones supremas de auditoría (ISA), destacan la necesidad de una gobernanza de datos eficaz para ayudar a los órganos de auditoría a seguir el ritmo de la digitalización del gobierno y la sociedad.[2] Las entidades públicas distintas de las ISA también están abordando los mismos problemas y desarrollando su propio marco de gobernanza de datos. Por ejemplo, en Nueva Zelanda, la agencia líder de datos de la administración (Stats NZ) ha desarrollado un marco de gobernanza de datos para el sector público que promueve una mejor gestión de datos y motiva a la administración pública a adoptar un «enfoque de ciclo de vida integral de datos» El marco motiva a los funcionarios públicos a pensar de manera más estratégica sobre la gobernanza, la gestión, la calidad y la responsabilidad de los datos que utilizan durante todo el ciclo de vida (es decir, desde el diseño y la fuente de los datos hasta su almacenamiento, publicación y eliminación) (OCDE, 2019[11]). En términos de calidad de los datos, hay varios principios rectores, entre ellos:

- *Relevancia*: la medida en que los datos satisfacen las necesidades de la organización y sus interesados legítimos.
- *Precisión y fiabilidad*: el grado en que los datos describen correcta y coherentemente el fenómeno que se examina.
- *Oportunidad y puntualidad*: la velocidad a la que se pueden obtener los datos y la fiabilidad de esta medición.
- *Accesibilidad y claridad*: la facilidad de acceso, la claridad y la asequibilidad de los datos disponibles.
- *Coherencia y comparabilidad*: la coherencia de los datos y la facilidad con la que se pueden combinar y comparar con otros datos.
- *Disponibilidad de metadatos*: la facilidad con la que se puede encontrar o comprender la información subyacente sobre los datos, su estructura y atributos (INTOSAI, 2019[12]).

El uso de datos de varias fuentes, que se preparan de forma independiente entre sí, puede generar una serie de desafíos para los organismos de control cuando se aplica a la detección del riesgo de fraude. La IGAE administra la BDNS y la utiliza para su propio análisis de riesgos, pero no es responsable de introducir los datos en la BDNS. Los organismos públicos, la Administración Local, la administración de Comunidades Autónomas, las fundaciones del sector público, entre otros, están obligadas a facilitar información a la BDNS. La IGAE no realiza evaluaciones de fiabilidad de datos en todos los datos. Como consumidor de datos, algunos de los problemas de calidad de los datos que son evidentes en los datos que usa la IGAE, como errores o valores omitidos, son responsabilidad de la agencia que introduce los datos. No obstante, los organismos de auditoría y control tienen la obligación de verificar la fiabilidad y validez de los datos de acuerdo con las normas internacionales, como las del Consejo de Normas Internacionales de Auditoría y Aseguramiento (IAASB) o el Comité de la Organización Patrocinadora de la Comisión Treadway (COSO). Además, las propias normas españolas para la obtención de pruebas de auditoría, como la Norma Internacional de Auditoría 500,3 enfatizan la necesidad de que los auditores evalúen la fiabilidad, exactitud e integridad de los datos. Por tanto, aunque la IGAE puede depender en cierta medida de la gobernanza, la gestión y los controles de calidad de los datos de los productores de datos (es decir, administraciones públicas u otras instituciones), también debe tomar medidas para evaluar de forma independiente los datos que obtiene.

Como se ilustra en el Capítulo 2, interpretar y depurar los datos para mejorar el modelo de riesgo de fraude de la IGAE requeriría mucho tiempo y recursos. Durante el transcurso del proceso sobre el que versa este informe, se hicieron evidentes las mejoras de «rápida ganancia» en la gestión de datos de la IGAE, como poder disponer un diccionario de datos que describa claramente los elementos de datos, o garantizar que los identificadores únicos se apliquen uniformemente en todos los conjuntos de datos. En

general, los datos de mala calidad pueden reflejar problemas como omisión de observaciones, información incorrecta o incorrectamente denominadas. Cualquiera de estos asuntos podría dificultar que los órganos de auditoría o control realicen análisis significativos y precisos de los riesgos y controles. Por ejemplo, en el contexto de la IGAE, los valores omitidos en los datos, aunque frecuentes, han sido un problema importante identificado al trabajar con varias bases de datos para desarrollar el modelo de riesgo. La información o los puntos de datos que faltan pueden reflejar errores o una supervisión laxa por parte de la entidad que introdujo los datos, pero también pueden deberse a una omisión intencionada. La implantación de verificaciones y controles para evitar que esto ocurra también puede ser un medio adicional para detectar y prevenir el fraude. Desde una perspectiva metodológica, depender de datos de mala calidad podría conducir a un muestreo ineficaz, por ejemplo, lo que significa que una serie de casos de fraude de subvenciones podrían pasar desapercibidos cada año. Los datos inexactos o incompletos también podrían sesgar negativamente técnicas más avanzadas, como el enfoque de aprendizaje automático elaborado en el Capítulo 2, lo que da como resultado modelos con poco poder predictivo y en última instancia una asignación ineficiente de los recursos de la IGAE.

La IGAE podría tomar medidas adicionales para garantizar que los datos de los sistemas y fuentes que utiliza sean fiables. Puesto en contexto, confirmar que los datos son fiables significa que la IGAE los consideraría suficientes y adecuados específicamente para el análisis de riesgos de fraude y la metodología que selecciona. En otras palabras, ¿los datos están completos, son exactos y describen realmente los conceptos claves que se están analizando? Como consumidora de datos, la IGAE podría trabajar con las instituciones y organizaciones fuente de datos en los que se basa para abordar algunos de los problemas indicados anteriormente y en el Capítulo 2, y garantizar la existencia de controles internos sólidos sobre los datos. Esto incluye las políticas y los procedimientos que rigen la recopilación, la gestión, el almacenamiento y el uso de datos.

En general, estos controles pueden clasificarse de tres formas: 1) Controles generales, 2) controles de aplicación y 3) controles de usuario (United States Government Accountability Office, 2019[13]) Los controles generales se aplican a los sistemas de información de la institución en su conjunto, mientras que los controles de aplicación son aquellos integrados en una aplicación en concreto para garantizar que todas las acciones dentro de ella sean válidas, precisas y completas. Los controles de usuario son aquellos administrados por personas para mejorar la fiabilidad del sistema de información. Al entender los controles que ya existen, la IGAE puede tener una mayor seguridad con respecto a la fiabilidad de los datos específicamente para evaluar los riesgos de fraude. Además, basándose en la experiencia de la OCDE al trabajar con los datos para detectar el fraude, la IGAE puede prestar especial atención a los siguientes problemas al juzgar la fiabilidad de los datos que utiliza:

- Verificar el número total de registros facilitados con las estadísticas de resumen.
- Comprobar si faltan observaciones, teniendo en cuenta todas las columnas o filas necesarias.
- Confirmar que ninguno de los registros esté duplicado.
- Buscar fechas fuera del rango deseado.
- Buscar valores que sean extremos atípicos.

La IGAE también puede consultar la documentación o manuales que expliquen cómo se diseñan los sistemas de información, pero en este caso también necesitaría verificar que la forma en que el sistema funciona se adhiere verdaderamente a este criterio. Como otra verificación adicional, los datos también podrían rastrearse hasta su origen para garantizar que los dos sean coherentes (United States Government Accountability Office, 2019[13]).

Desarrollar capacidad para análisis y evaluaciones de riesgos basadas en datos, en concreto, competencias para trabajar con conjuntos de datos y visualización de datos a gran escala.

Arquitectura de datos, infraestructura de datos y capacidad de implantación, han sido destacados por funcionarios de la IGAE como algunas de sus principales prioridades para mejorar el uso de datos y análisis en general. Estas áreas han sido el núcleo de varias recomendaciones de la OCDE para la IGAE y la ONA, para fortalecer el sistema de supervisión continua, en parte mediante la automatización de los procesos de importación de datos, así como mejorando los esfuerzos para validar y corroborar los datos autoinformados (OCDE, 2021[14]). En el contexto de la evaluación de riesgos de fraude, dado el almacenamiento y la escala de la mayoría de las subvenciones y conjuntos de datos relacionados que utiliza la IGAE, o a los que podría acceder en el futuro, como datos del registro mercantil, es fundamental para la extracción de datos la capacidad de los servidores de la administración pública para gestionar el volumen de datos de manera oportuna y fiable. Grandes conjuntos de datos de varios millones de registros, incluso la limpieza de datos básica y el trabajo analítico, pueden requerir el uso de servidores de gran capacidad. Los funcionarios de la IGAE destacaron la necesidad de mejorar la infraestructura de datos de la IGAE. Sin embargo, para los propósitos de este proyecto y para evaluar los riesgos de fraude en los datos de las subvenciones públicas, la infraestructura existente es suficiente para formas más avanzadas de análisis de riesgos, como lo demuestra la metodología de aprendizaje automático descrita en el Capítulo 2.

Como una necesidad más inmediata de implementar dicha metodología y análisis similares, la IGAE puede desarrollar sus competencias digitales internas para manejar conjuntos de datos a gran escala (es decir, cientos de miles o millones de observaciones) e implementar métodos estadísticos avanzados, como *Random Forests*, como se describe en el Capítulo 2. La fase de preprocesado - creación, extracción, fusión y organización de los conjuntos de datos antes del análisis real - consume mucho tiempo, es costosa y requiere conocimientos básicos de datos para tratar y limpiar los datos. Los costes a menudo dependen de la calidad y transparencia de los sistemas de datos públicos. Salvo algunas excepciones, la IGAE tiene capacidad para acceder a muchas bases de datos que pueden usarse para la detección de fraude, pero consumir tiempo procesando datos de poca calidad puede aumentar los costes.

Además de la calidad de los datos, los generadores de costes pueden incluir la creación de conjuntos de datos de subvenciones digitalizados, centralizados y estructurados, así como el formato para almacenarlos y la correspondiente facilidad para extraer los campos relevantes. Para este proyecto, la OCDE ha apoyado a la IGAE en la creación de una base de datos que pueda utilizarse para el análisis de riesgos de fraude, independientemente de la metodología utilizada, reduciendo así dichos costes en el futuro. Sin embargo, los datos, como los riesgos en sí mismos, no son estáticos y requieren la combinación adecuada de conocimientos técnicos y experiencia en riesgos para actualizarse continuamente. Por ejemplo, para mejorar aún más su capacidad para realizar evaluaciones de riesgo de fraude basadas en datos, la IGAE puede formar un equipo multidisciplinar con experiencia en operaciones de subvenciones, gestión de riesgos de fraude, análisis y visualización de datos.

La metodología del Capítulo 2 ha utilizado *software* de código abierto (es decir, Python y R). Si bien muchas instituciones de auditoría dependen de *software* de pago (como IDEA, ACL, SAS o Stata), no existe una solución única para todas, y muchas entidades han desarrollado análisis eficaces con herramientas de código abierto, en busca de una herramienta más sólida que Excel. En general, los objetivos del análisis, así como las aptitudes y la experiencia de los auditores, determinarán qué herramienta es la más apropiada. Por ejemplo, el Tribunal de Cuentas de Austria (ACA) desarrolló una herramienta para monitorizar la salud financiera de los municipios austriacos. La herramienta funciona principalmente a través del *software* estadístico R y permite comparar en función de diversos criterios a los municipios, e identificar aquellos que presentan el riesgo financiero más elevado. La ACA percibió que el *software* R era más apropiado que Excel para analizar *big data*, menos propenso a errores, y el código

en R podían reutilizarse fácilmente en evaluaciones futuras con pocas adaptaciones. La curva de aprendizaje para los analistas del ACA fue significativa, según los funcionarios del ACA, dado el nivel de experiencia técnica detallada requerida. No obstante, tener experiencia interna en estas aplicaciones y lenguajes de codificación se ha convertido en un conjunto de aptitudes estándar para muchas instituciones de auditoría que han avanzado en sus capacidades analíticas en los últimos años.

La capacidad de aprovechar los datos y el análisis va de la mano de aptitudes de visualización de datos. Visualizar datos de manera que ayude a los usuarios a comprender y actuar sobre los resultados requiere conocimiento de los principios de visualización de datos, así como familiaridad, si no experiencia, con *software* especializado que pueda generar paneles de control y facilite la comprensión de los riesgos por parte de los auditores (por ejemplo, paquete R Shiny, o Tableau). Los funcionarios de la IGAE destacaron la necesidad de dichas herramientas y paneles de control para respaldar los análisis de la BDNS como una de sus principales prioridades y necesidades. Actualmente, la IGAE hace poco uso de visualizaciones de datos para evaluar grandes riesgos de fraude. Los análisis de redes para identificar conflictos de interés son un área que se presta bien para visualizar riesgos (ver Capítulo 2).

Los usuarios que tienen un profundo conocimiento sobre los procesos de subvenciones, las bases de datos disponibles y los riesgos son fundamentales para crear una capacidad analítica eficaz y un enfoque basado en datos para evaluar riesgos. La IGAE tiene un equipo con una base sólida en todas estas áreas, pero podría invertir más en adquirir experiencia en análisis y visualización de datos para avanzar en sus capacidades digitales. La creación, las pruebas de validación y el análisis de modelos de riesgo de fraude exigen un profundo conocimiento del proceso de concesión y ejecución de subvenciones, así como aptitudes analíticas avanzadas. El conocimiento específico sobre ayudas y subvenciones ayuda a comprender el alcance de los datos y las definiciones de variables, así como el marco normativo que rige el ciclo de subvenciones. Estos diversos problemas de capacidad, muchos de los cuales reflejan las necesidades de la IGAE, destacan la importancia de tener objetivos y prioridades claros al desarrollar una capacidad analítica.

Si bien los nuevos enfoques basados en datos pueden ser un catalizador para un cambio más amplio, hacer un uso eficaz de los datos y el análisis requiere más que simplemente introducir nuevas herramientas, técnicas o fuentes de datos. Además, es probable que las cuestiones sobre la creación de una capacidad analítica deban tener en cuenta otros aspectos del trabajo de la IGAE más allá del alcance de este proyecto. Por ejemplo, la forma en que la IGAE desarrolla su capacidad para mejorar sus análisis para evaluar riesgos de fraude probablemente puede relacionarse con una estrategia de digitalización más amplia, metas y recursos para mejorar la arquitectura e infraestructura de datos, u objetivos institucionales para actividades de control más específicas y eficaces. Recuadro 1.2 describe la experiencia de la estrategia del Servicio de Auditoría Interna de la Unión Europea para mejorar su función analítica adoptando un enfoque institucional.

Recuadro 1.2. Desarrollo de una estrategia de analítica en el Servicio de Auditoría Interna de la Unión Europea

El Servicio de Auditoría Interna (IAS) de la Unión Europea han avanzado a pasos agigantados en el uso de análisis y tecnología en sus investigaciones y auditorías en los últimos años. Esto se consiguió, desde el principio, ideando e implantando una estrategia de análisis sólida y consistente. De entrada, el equipo de Informática existente realizó un análisis extenso de áreas de mejora, incluidas innovaciones y nuevas tecnologías que el servicio podría incorporar a su trabajo. Los IAS también establecieron un grupo interno para continuar este trabajo, incluido el descubrimiento de formas en que los datos y la tecnología podrían usarse en actividades novedosas, para mantenerse al día de las buenas prácticas actuales y que el departamento sea más eficiente a través de la analítica. Para impulsar este esfuerzo, el IAS diseñó una estrategia a largo plazo en torno a la analítica centrada en tres áreas claves: 1) desarrollar un inventario sólido de conocimientos y aptitudes, 2) iniciar proyectos piloto y 3) compartir conocimientos. Crear una estrategia única para toda la organización ayudó al IAS a planificar las auditorías de manera más eficaz, entre otras ventajas.

Fuente: (Barrigon, 2020[15])

Tener consciencia las trampas relacionadas con los indicadores compuestos de riesgo, así como de los sesgos

Si bien el control basado en el riesgo es parte del plan anual de la IGAE, la selección de auditorías e investigaciones basadas en riesgos percibidos tiene como objetivo, en última instancia, optimizar la relación calidad-coste en el uso del dinero de los contribuyentes. Por tanto, es fundamental que la IGAE sea consciente de algunas de las dificultades inherentes a los enfoques típicos de las evaluaciones de riesgos, y que reduzca el riesgo de falsos positivos y falsos negativos. Una de las formas más frecuentes de crear indicadores de riesgo (compuestos) se basa en la selección manual de características observadas de casos conocidos destacados y generalizarlos, aplicando los mismos indicadores a todo el conjunto de datos de casos.

Este enfoque adolece de dos grandes inconvenientes. En primer lugar, provoca el llamado sesgo de selección, lo que significa que se han considerado casos particulares, suponiendo que sus características son generalizables a otras observaciones, sin que exista prueba de que sean típicas o representativas de todo tipo de esquemas de fraude. En segundo lugar, estos enfoques no suelen tener en cuenta la prevalencia de indicadores de riesgo seleccionados (o señales de alerta) entre los casos limpios y desconocidos. En otras palabras, suelen producir altos índices de falsos positivos, lo que significa que a menudo señalan riesgos de fraude cuando no hay fraude. En tercer lugar, por lo general, estos enfoques aplican una media sencilla de las señales de alerta individuales para producir una puntuación compuesta, ya que no entienden cómo los diferentes indicadores concurren entre sí, o cuáles son más importantes.

Aunque no es el único enfoque, se seleccionó la metodología descrita en el Capítulo 2 porque aborda estas deficiencias, y como se analiza a continuación, permite a la IGAE solucionar algunas de las peculiaridades de los datos que usa. El método de aprendizaje automático del Capítulo 2 se generaliza a partir de todos los casos probados en el pasado (es decir, casos sancionados) para identificar qué factores influyeron en la probabilidad de ser sancionado. Este enfoque conduce a una calificación única de riesgo compuesta por todas las características relevantes en los datos, con ponderaciones de cada característica definidas para optimizar el poder predictivo. El enfoque también aborda explícitamente el problema de los falsos positivos y los falsos negativos, aprendiendo tanto de los casos positivos probados (sancionados) como de los probables negativos (no sancionados). No obstante, ninguna metodología está

completamente libre del riesgo de sesgos o imprecisiones. Sin embargo, tener en cuenta estos y las tendencias inherentes de metodologías específicas relacionadas con estos temas puede ayudar a la IGAE a adoptar una visión fundamentada para fortalecer su metodología actual de evaluación de riesgos. Recuadro 1.3 explora más a fondo cómo la IGAE puede controlar los sesgos en sus modelos, basándose en las principales prácticas internacionales.

Recuadro 1.3. Abordar los sesgos en los modelos de aprendizaje automático

Los modelos de aprendizaje automático se entrenan en función de los datos disponibles, por lo que pueden estar intrínsecamente sesgados. El diseñador del algoritmo también puede amplificar aún más estos sesgos, intencional o inconscientemente. Esto es de especial interés para auditores o investigadores de fraude, para quienes la objetividad es de suma importancia. Varias instituciones, incluidos los organismos de auditoría y los grupos de expertos (por ejemplo, la Brookings Institution), han publicado unas orientaciones sobre cómo auditar la inteligencia artificial y cómo verificar los sesgos en los algoritmos, para mejorar el modelo de aprendizaje automático. Estas recogen, entre otros, los siguientes puntos:

- Los algoritmos se pueden auditar de forma periódica e independiente. La auditoría debería incluir evaluar el proceso de recopilación de datos, monitorizar cómo funciona el programa y verificar si están tratando de manera justa los subgrupos sensibles.
- El programa debería compararse con evaluaciones de riesgo preparadas por humanos para ver si en realidad es más eficaz.
- Se puede verificar si los algoritmos cumplen las leyes de no discriminación.
- Los operadores de algoritmos pueden intentar aumentar la interacción humana con el programa, esforzándose por garantizar que se entienda el código y las métricas que se utilizan y que se tenga en cuenta su relación con las principales desigualdades sociales.
- La entidad que opera los algoritmos debería considerar redactar una declaración de impacto de sesgo formal, para documentar su consideración y estrategia consciente al gestionar este desafío.

Según la Brookings Institution, algunas preguntas que se pueden tener en cuenta e incluir en una declaración de este tipo incluyen:

- ¿Qué hará la decisión automatizada?
- ¿Sobre qué audiencia trabaja el algoritmo y quiénes se verán más afectados por él?
- ¿Tiene la organización datos de formación para hacer las predicciones correctas sobre la decisión?
- ¿Son los datos de entrenamiento suficientemente diversos y fiables? ¿Cuál es el ciclo de vida de los datos del algoritmo?
- ¿Qué grupos pueden ser tratados injustamente o pueden verse afectados de manera desproporcionada por los procesos de entrenamiento del modelo y el consiguiente análisis?
- ¿Cómo se detectarán los posibles sesgos?
- ¿Cómo y cuándo se probará el algoritmo? ¿A quiénes estarán dirigidas las pruebas?
- ¿Cuál será el umbral para medir y corregir el sesgo en el algoritmo?
- ¿Cuáles son los incentivos del operador del algoritmo?
- ¿Qué ganancias se esperan con el desarrollo del algoritmo?
- ¿Cuáles son los posibles malos resultados y cómo se dará cuenta la organización de ellos?

- ¿Cómo de abierto (por ejemplo, en código o intención) será el proceso de diseño del algoritmo para las partes interesadas internas y externas?
- ¿Cómo se intervendrá si se predice que podría haber malos resultados asociados al desarrollo o implantación del algoritmo?
- ¿Cómo se están implicando otras partes interesadas?
- ¿Cuál es el ciclo de retroalimentación del algoritmo para desarrolladores, usuarios y partes interesadas?
- ¿Tienen las organizaciones de la sociedad civil algún papel en el diseño del algoritmo?
- ¿Se ha tenido en cuenta la diversidad en el diseño y ejecución?
- ¿Tendrá el algoritmo implicaciones para los grupos culturales y se desarrollará de manera diferente en contextos culturales?
- ¿Es el equipo de diseño lo suficientemente representativo como para capturar estos matices y predecir la aplicación del algoritmo en diferentes contextos culturales? Si no es así, ¿qué medidas se están tomando para hacer que estos escenarios sean más destacados y comprensibles para los diseñadores?
- Teniendo en cuenta el propósito del algoritmo, ¿los datos de entrenamiento son lo suficientemente diversos?
- ¿Existen restricciones legales que las organizaciones deberían revisar para asegurarse de que el algoritmo sea tanto legal como ético?

Fuente: (Canadian Audit and Accountability Foundation, 2019[16]); (Lee, Resnick and Barton, 2019[17])

Conclusión

La IGAE ha desarrollado una base sólida para avanzar en el uso de datos y análisis para evaluar los riesgos de fraude en los datos de subvenciones públicas. Las habiidades y conocimientos que posee internamente, en concreto con respecto a los procesos de concesión de subvenciones, los riesgos existentes y las complejidades de las bases de datos pertinentes, son elementos claves de la capacidad y la experiencia necesarias para evaluar eficazmente los riesgos de fraude en las subvenciones. No existe una herramienta o método analítico que pueda reemplazar este conocimiento o juicio de expertos. Además, para algunos autores, las comprobaciones sobre el terreno y los mecanismos internos de denuncia de fraude se perciben como la medida de detección de fraude más eficaz y están mejor posicionadas que el análisis de datos o la minería de datos (Dozhdeva and Mendez, 2020[18]). No obstante, con una administración pública y una sociedad cada vez más digitales, los órganos de supervisión como la IGAE tendrán que evolucionar por necesidad y no por elección.

Sobre la base de su sólida experiencia y conocimientos, la IGAE puede valorar incorporar capacidades para aprovechar todo el potencial de las bases de datos existentes a su disposición, en concreto fortalecer su capacidad para trabajar con grandes conjuntos de datos y la visualización de datos. Al mismo tiempo, la IGAE puede seguir mejorando su gestión de datos y verificar la calidad de los datos, para facilitar la fusión de conjuntos de datos y evaluar el riesgo de fraude en subvenciones públicas. Se trata de acciones que ayudarían a la IGAE a madurar desde una perspectiva analítica, independientemente de que decida adoptar la metodología específica del Capítulo 2. Avanzar en el uso de datos y análisis ayudaría a la IGAE no solo a recopilar información más predictiva sobre los riesgos en los programas de subvenciones públicas, sino también a ser más eficiente y eficaz en el uso del dinero de los contribuyentes.

Referencias

Barrigon, F. (2020), "Innovation and digital auditing – the journey of the European Commission's IAS towards state-of-the-art technologies", *ECA Journal*, Vol. 1/2020, pp. 97-101, https://www.eca.europa.eu/Lists/ECADocuments/JOURNAL20_01/JOURNAL20_01.pdf. [15]

Canadian Audit and Accountability Foundation (2019), *Artificial Intelligence and Auditing: Overview of Potential Impact on Public Sector Auditors*, https://caaf-fcar.ca/en/performance-audit/research-and-methodology/research-highlights/3455-research-highlights-3. [16]

Centers for Medicare & Medicaid Services (2014), *Report to Congress, Fraud Prevention System, Second Implementation Year*, https://www.cms.gov/About-CMS/Components/CPI/Widgets/Fraud_Prevention_System_2ndYear.pdf (accessed on 13 August 2021). [10]

Dozhdeva, V. and C. Mendez (2020), *Is fraud risk management in cohesion policy effective and proporitionate?*, https://www.eprc-strath.eu/public/dam/jcr:dbcbcfde-e024-44a0-a11b-b12456ffe0c5/EPRP%20121%20-%20IQ_Net_Thematic%20paper%2047(2).pdf. [18]

European Commission Anti-Fraud Office (OLAF) (2017), *Handbook on Reporting on Irregularities in Shared Management*, https://www.eu-skladi.si/sl/dokumenti/navodila/handbook-irregularity-reporting-final.pdf (accessed on 13 August 2021). [19]

Gobierno de España (2019), *Real Decreto 130/2019*, https://www.boe.es/eli/es/rd/2019/03/08/130 (accessed on 13 August 2021). [6]

Gobierno de España (2003), *Ley 38/2003, de 17 de noviembre, General de Subvenciones*, https://www.boe.es/buscar/pdf/2003/BOE-A-2003-20977-consolidado.pdf (accessed on 13 August 2021). [4]

Gobierno de España (2003), *Ley 47/2003, de 26 de noviembre, Presupuestaria*, https://www.boe.es/buscar/act.php?id=BOE-A-2003-21614&p=20201231&tn=6. [3]

IGAE (2020), *Activity report 2019 (Memoria de actividades 2019)*, https://www.igae.pap.hacienda.gob.es/sitios/igae/es-ES/QuienesSomos/Documents/Memoria_2019.pdf. [2]

IGAE (2020), *Approval Of The Audit And Financial Control Plan Of Subsidies 2021 (Aprueban El Plan De Auditorías Y Control Financiero De Subvenciones 2021)*, https://www.igae.pap.hacienda.gob.es/sitios/igae/es-ES/Control/CFPyAP/Documents/Resoluci%C3%B3n%20Plan%20Auditor%C3%ADa%20Pbca%20y%20CFP%202021.pdf (accessed on 13 August 2021). [5]

IGAE Ministerio de Hacienda, (2021), *National System of Publicity for Subsidies and Public Aid (Sistema Nacional de Publicidad de Subvenciones y Ayudas Públicas)*. [9]

INTOSAI (2019), *Training Tool on Environmental Data: Resources and Options for Supreme Audit Institutions*, https://www.environmental-auditing.org/media/113693/23g-wgea_environmental-data_2019-fin.pdf (accessed on 13 August 2021). [12]

Lee, N., P. Resnick and G. Barton (2019), “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms”, *Brookings Institute*, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ (accessed on 13 August 2021). [17]

OCDE (2021), *Mejora de la Responsabilidad Pública en España Mediante la Supervisión Continua*, Estudios de la OCDE sobre Gobernanza Pública, OECD Publishing, Paris, https://doi.org/10.1787/4962ce0f-es. [14]

OCDE (2019), *The Path to Becoming a Data-Driven Public Sector*, OECD Digital Government Studies, OECD Publishing, Paris, https://dx.doi.org/10.1787/059814a7-en. [11]

OCDE (2014), *Spain: From Administrative Reform to Continuous Improvement*, OECD Public Governance Reviews, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264210592-en. [1]

Unión Europea (2014), *Commission Regulation (EU) No 651/2014*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02014R0651-20210405 (accessed on 13 August 2021). [7]

Unión Europea (2014), *Commission Regulation (EU) No 702/2014*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02014R0702-20201210 (accessed on 13 August 2021). [8]

United States Government Accountability Office (2019), *Assessing Data Reliability*, https://www.gao.gov/assets/gao-20-283g.pdf (accessed on 13 August 2021). [13]

Notas

[1] La Unión Europea define las irregularidades como "toda infracción de una disposición del Derecho comunitario resultante de una acción u omisión de un agente económico que tenga o pueda tener por efecto a un perjuicio para el presupuesto general de las Comunidades o para los presupuestos administrados por éstas, ya sea por la disminución o la pérdida de ingresos procedentes de recursos propios percibidos directamente por cuenta de las Comunidades, o ya sea por un gasto no justificado". Por otra parte, se considera fraude "en materia de gastos, todo acto u omisión intencional relativo a la utilización o a la presentación de declaraciones o de documentos falsos, inexactos o incompletos, que tengan por efecto la desviación o la retención indebida de fondos del presupuesto general de la UE o de los presupuestos administrados por ésta o en su nombre, o la no comunicación de información en violación de una obligación específica, con el mismo efecto, o la utilización indebida de dichos fondos para fines distintos de aquellos para los que fueron concedidos inicialmente." (European Commission Anti-Fraud Office (OLAF), 2017[19]).

[2] Por ejemplo, consulte el informe de investigación de la Organización de Instituciones Supremas de Auditoría de África sobre la integración de macrodatos en la auditoría del sector público (https://afrosai-e.org.za/wp-content/uploads/2020/12/Research-Paper-Integrating-Big-Data-in-Public-Sector-Auditing.pdf); la herramienta de formación sobre datos medioambientales publicada por el Grupo de Trabajo de Auditoría Medioambiental de la INTOSAI (https://www.environmental-auditing.org/media/113693/23g-wgea_environmental-data_2019-fin.pdf); o las experiencias del Tribunal

de Cuentas de los Países Bajos en el desarrollo de un marco de auditoría para algoritmos (http://intosaijournal.org/developing-an-audit-framework-for-algorithms/).

[3] La Norma Internacional de Auditoría 500 fue adaptada de las Normas Internacionales de Auditoría emitidas por la Federación Internacional de Cuentas a través del IAASB.

2 Fraude en subvenciones públicas: pilotar un modelo de riesgo basado en datos en España

Este capítulo presenta una prueba de concepto para un modelo de riesgo que la Intervención General de la Administración del Estado (IGAE) de España puede emplear para evaluar los riesgos de fraude y detectar posibles casos de fraude. El capítulo presenta una descripción general de la metodología de aprendizaje automático que subyace en el modelo de riesgo, así como un relato detallado de cómo se construyó el modelo, basado en datos que están fácilmente disponibles para la IGAE. El capítulo concluye con una exposición de los resultados del modelo y recomendaciones para que la IGAE se base en la prueba de concepto.

Introducción

Los marcos de evaluación de riesgos de fraude basados en datos pueden tener múltiples usos, entre los cuales es fundamental identificar las prioridades de investigación. Cuando los recursos de investigación son escasos, y es probable que una selección aleatoria de casos para la investigación arroje una tasa de éxito baja (por ejemplo, porque el fraude es poco frecuente en la población objetivo), una selección de casos basada en riesgos puede reportar beneficios significativos. Con este fin, una calificación de riesgo asignada a todos los casos potencialmente investigados puede contribuir a priorizar los casos que deben investigarse. Por lo general, esto no implica una automatización completa de la selección de casos, pero sí ofrece una aportación crucial en el proceso de toma de decisiones de la organización.

Para que una evaluación de riesgos a gran escala reporte beneficios, debe ser lo suficientemente precisa como para ser utilizada para operaciones u organizaciones de calificación de riesgos de manera continua, incluidos los casos nuevos. En general, las calificaciones de riesgo pueden definirse como válidas para dichos propósitos, ya sea definiendo explícitamente los factores de riesgo a partir de relaciones conocidas y descripciones de riesgo (por ejemplo, el propietario último de la organización receptora de la subvención reside en un paraíso fiscal), o definiendo la combinación de características de riesgo a través de medios estadísticos, incluido el aprendizaje automático de investigaciones anteriores. De cualquier manera, lo que es crucial es que el modelo de riesgo no solo tenga en cuenta los casos fraudulentos conocidos y sus características, sino que también tenga en cuenta las características de un grupo mucho mayor de casos que no hayan sido investigados, por lo que se desconoce su situación respecto al fraude. En resumen, la validación de indicadores y la mejora continua son de crucial importancia, como se muestra en este capítulo.

Cuando se desarrollan nuevos enfoques analíticos, la comprensión y el aprendizaje suelen derivarse de la práctica. Este es el motivo por el que muchas instituciones de auditoría, por ejemplo, han creado «Laboratorios de innovación» y comunidades internas de práctica para probar y experimentar con nuevas técnicas de auditoría, análisis y tecnologías y uso de los datos. Este enfoque incremental permite a los organismos de auditoría y control asumir riesgos medidos y contener los costes, antes de ampliar o reducir las iniciativas piloto. Con este espíritu y en respuesta al interés de la Intervención General de la Administración del Estado (IGAE) en fortalecer su uso de datos para detectar riesgos de fraude en subvenciones, este capítulo presenta una prueba de concepto para un modelo de riesgo basado en datos para que la IGAE lo adopte en parte o en su totalidad.

La metodología de este capítulo tiene como objetivo hacer uso de los datos que ya estaban a disposición de la IGAE, incluida la Base de Datos Nacional de Subvenciones (BDNS) y, al hacerlo, implícitamente tiene en cuenta el contexto de la IGAE. Como se indica en el Capítulo 1, como cualquier inversión para mejorar la gobernanza de datos, la gestión de datos o el análisis, este enfoque puede requerir inversiones en aptitudes y habilidades digitales. Por este motivo, el capítulo proporciona una descripción detallada de todas las etapas de la metodología y su desarrollo para respaldar la propia evaluación de la IGAE de lo que es capaz de hacer con sus recursos y habilidades internas existentes. Además, el proceso de desarrollo de la prueba de concepto para el modelo de riesgo ha derivado en diversos descubrimientos y en la identificación de áreas de mejora, que se abordan en la sección de resultados.

Aspectos generales del modelo de aprendizaje automático

Una breve introducción al aprendizaje automático para las evaluaciones de riesgos

El enfoque actual de la IGAE para evaluar riesgos de fraude, detallado en el Capítulo 1, se describe en su Plan de Auditorías y Control Financiero de Subvenciones y Ayudas Públicas para 2021. La IGAE tiene en cuenta el importe de la subvención, los niveles previos de fraude y otros factores cualitativos, como los

procedimientos de justificación y verificación. El modelo de aprendizaje automático descrito en este capítulo avanza en un enfoque más basado en datos, que puede complementar los procesos existentes de la IGAE. En realidad, dadas las limitaciones de recursos, la IGAE solo puede realizar un número finito de actividades de control en un año. La metodología de aprendizaje automático descrita en este capítulo no debe sustituir el criterio del auditor. Por ejemplo, el modelo puede detectar casos de posible fraude, pero el auditor también necesitará evaluar cuál de estos casos es el más rentable en términos de actividades de investigación o control adicionales. Teniendo en cuenta este matiz, el modelo puede ser una aportación útil para las decisiones de los auditores y ayudar a la IGAE a dirigir sus recursos de manera más eficaz.

El modelo de riesgo desarrollado para apoyar la IGAE se basa en una metodología de *random forest*. Los *random forests* son un método de aprendizaje automático supervisado que predice el resultado mediante la formación de varios árboles de decisión con características determinadas (Breiman, 2001[1]). Es especialmente adecuado para conjuntos de datos con una gran cantidad de variables explicativas o indicadores de riesgo potencial. Al utilizar *random forests*, es posible incluir una amplia lista de factores explicativos de diferentes tipos (numéricos y categóricos).

Selección de la metodología

Para analizar los datos utilizando métodos de aprendizaje automático como *random forests*, el conjunto de datos se limpia previamente, eliminando los valores omitidos y las variables que carecen de varianza (es decir, donde las variables toman casi siempre el mismo valor en todo el conjunto de datos). Los *random forests* permiten trabajar con una gran cantidad de observaciones y variables, realizar entrenamiento de algoritmos en una muestra reducida y equilibrada, y probar modelos en una muestra reservada. Los algoritmos de *Random Forest* son sensibles a los valores omitidos. Por esta razón, se descartaron las variables con altos índices de omisión. El método también es sensible al desequilibrio en la variable dependiente (es decir, sancionado frente a no sancionado), como se describe a continuación. En general, el enfoque se puede dividir en los siguientes pasos:

1. Identificar qué beneficiarios fueron sancionados y luego marcar todas las concesiones a las organizaciones sancionadas en los últimos 2 ó 3 años antes de las sanciones. En este período, es muy probable que se haya producido una actividad fraudulenta probada. Esto da un conjunto completo de casos positivos probados (concesiones sancionadas); sin embargo, deja una muestra muy grande de casos sin etiquetar (no sancionados). Algunos de estos casos probablemente deberían haber sido sancionados, pero no fueron investigados, y otros son casos negativos reales en los que no habría habido sanción incluso si se hubieran investigado. En otras palabras, el conjunto de datos está muy desequilibrado. En la mayoría de los casos, se desconoce si la concesión no fue sancionada porque no fue investigada o porque no se descubrieron infracciones. Por tanto, la mayoría de las observaciones no son positivas ni negativas, sino que no están etiquetadas.
2. Elegir el método que se adapte al problema particular de los datos, es decir, una muestra desequilibrada y la presencia de una submuestra grande sin etiquetar. Para estos fines, se aplica un modelo de insaculación (*bagging)* positivo sin etiquetar (PU). Este método de aprendizaje automático permite entrenar un modelo en muestras aleatorias de observaciones, tanto positivas como sin etiquetar, para asignar un estado probablemente negativo (no sancionado) y un estado probablemente positivo (sancionado) a los casos sin etiquetar. El Recuadro 2.1 proporciona información adicional sobre la insaculación (*bagging)* de PU y los modelos de *random forest.*
3. Una vez asignadas las etiquetas, utilizar el conjunto de datos reetiquetado para entrenar al modelo, e identificar los factores que influyen en la probabilidad de ser sancionado. La influencia puede ser tanto positiva como negativa. Luego, el modelo calcula la probabilidad de que cada concesión sea sancionada por cualquier número de observaciones.

4. Una vez que el modelo esté entrenado y logre una precisión suficiente, aplicarlo al conjunto de datos completo de concesiones, para predecir una calificación de riesgo de fraude para todas las observaciones.[1]

Recuadro 2.1. Descripción general del aprendizaje y la insaculación de positivos sin etiquetar

El aprendizaje positivo sin etiquetar (PU) es una técnica de aprendizaje automático semisupervisada que permite trabajar con datos muy desequilibrados (Elkan and Noto, 2008[2]). El aprendizaje PU puede utilizarse en casos en los que la mayoría de las observaciones disponibles pertenecen a casos sin etiquetar. Esto incluye situaciones en las que una variable binaria (es decir, valores de 1 y 0) tiene observaciones positivas (1) que aparecen solo en caso de tratamiento, y cuando se desconoce si los casos negativos restantes (0) fueron tratados, pero siguieron siendo negativos, o no fueron tratados de ninguna manera. El aprendizaje PU observa todos los casos positivos y negativos, identifica las características más típicas de cada uno, y vuelve a etiquetar las observaciones como corresponde.

Un enfoque de *insaculación* PU consta de varios pasos (Li and Hua, 2014[3]). Primero, implica construir un clasificador analizando la variedad y combinación de factores que influyen en los resultados positivos y negativos. Para construir un clasificador, se crea un subconjunto de datos, que consta de todos los casos positivos y una muestra aleatoria de los no etiquetados. Este clasificador se aplica además al resto de casos sin etiquetar, para asignar las puntuaciones de probabilidad para el resto de las observaciones. Cada paso se repite varias veces, y luego se calcula la puntuación media recibida por cada observación.

Después de volver a etiquetar todas las observaciones, se divide en muestras de entrenamiento y prueba. La proporción de la división es flexible, pero suele estar entre el 60 % y el 70 % para la muestra de entrenamiento, y entre el 30 % y el 40 % para la muestra de prueba. Después, se aplica el método de *random forests* al conjunto de datos de entrenamiento. Los parámetros del modelo se pueden especificar manualmente, incluido el número de árboles, el número máximo de características en cada árbol individual y el tamaño de los nodos terminales. La elección de los parámetros depende del tamaño total del conjunto de datos, es decir, el número de observaciones e indicadores incluidos en el modelo. Después de aplicar el método de *random forests* a la muestra de entrenamiento, se pueden predecir las probabilidades de salida para el resto de los datos.

Además, para identificar el impacto de cada indicador, los valores SHAP (explicaciones aditivas de Shapley) se pueden calcular una vez que se construye el modelo. Los valores SHAP muestran cuánto y en qué dirección (positiva o negativa) se ha modificado la salida prevista para un indicador determinado. Para estimar el ajuste del modelo, deben calcularse parámetros tales como exactitud, repetición y precisión. Mediante ellos se calcula el número de predicciones de calificación correctas en términos absolutos o relativos.

Fuente: (Mordelet and Vert, 2014[4])

Consideración de puntos fuertes, puntos débiles y supuestos

La validez del análisis depende de dos factores: la calidad del conjunto de datos de entrenamiento y la disponibilidad de las características relevantes de la concesión, la ayuda y el beneficiario. En primer lugar, el principal indicador que diferencia los casos fraudulentos de los no fraudulentos es la presencia de sanciones. Para que el aprendizaje positivo sin etiquetar genere resultados válidos, se ha supuesto que los casos positivos se seleccionaron al azar, por lo que son una muestra representativa de todos los casos

positivos. Esto también implica que si la muestra de sanciones observadas no recoge algunos esquemas típicos de fraude (es decir, ni siquiera se encuentra un ejemplo entre los casos de sanciones observados), el modelo de aprendizaje automático no capturará dichos tipos de fraude, y por tanto, estará sesgado. De manera similar, si los casos que se seleccionaron siguiendo una variable en particular – por ejemplo, el tamaño del beneficiario -, el modelo sobreestimará la importancia de dicha variable en la predicción del riesgo. En otras palabras, el aprendizaje automático supervisado utiliza la información proporcionada por casos probados. Por tanto, si hay un sesgo en la muestra de concesiones sancionadas, se replicará en el proceso de predicción.

En segundo lugar, el modelo de aprendizaje automático solo puede conocer las características del fraude que capturan los datos. La presencia de ciertos indicadores en el conjunto de datos influye en el poder predictivo del modelo: si faltan algunos indicadores cruciales en los datos, el modelo no los tendrá en cuenta. Las características o indicadores que faltan también implican que la lista final de indicadores influyentes puede estar sesgada, exagerando la importancia de aquellas características que están correlacionadas con características influyentes pero no observadas (por ejemplo, si se encuentra que una región en particular tiene un mayor riesgo, en realidad, puede significar que algunas entidades en esa región tienen características de riesgo, como vínculos con políticos corruptos, y no que la región misma, su cultura, estructuras administrativas, etc., sean más propensas al fraude). Sin embargo, el método de aprendizaje automático elegido basado en *random forests* es especialmente adecuado para grandes conjuntos de datos con una gran cantidad de variables explicativas o indicadores de riesgo potencial (James et al., 2015[5]) Es posible incluir una amplia lista de factores explicativos de diferentes tipos (numéricos y categóricos).

Desarrollo de una prueba de concepto para un modelo de riesgo basado en datos

Identificar fuentes de datos y variables relevantes para evaluar riesgos de fraude de subvenciones

Los datos proporcionados por la IGAE constan de 17 conjuntos de datos que abarcan diferentes bloques de información sobre concesiones, terceros, proyectos, subvenciones y beneficiarios. Podrían agruparse en tres categorías principales.

- La primera categoría consta de siete conjuntos de datos que abarcan información sobre la convocatoria, como ubicación, tipo de actividad económica, objetivos e instrumentos.[2]
- La segunda categoría abarca información sobre concesiones, incluida información sobre reintegros, proyectos, devoluciones y detalles de las concesiones a beneficiarios.[3]
- La tercera categoría incluye conjuntos de datos que abarcan información sobre los propios beneficiarios, que pueden incluir una variedad de actores responsables de implantar un proyecto (por ejemplo, una entidad pública, contratista o subcontratista), si un beneficiario fue sancionado o inhabilitado, así como el tipo de actividad económica, ubicación, etc.[4]

En total, estos conjuntos de datos constan de alrededor de 100 variables que cubren detalles de las concesiones (importe, fecha de resolución, tipo de actividad económica, etc.), convocatorias de ayudas (publicidad, tipo de apoyo económico, base normativa, etc.) y datos de terceros (ubicación, naturaleza jurídica, actividades económicas, etc.). El período cubierto es 2018-2020.

Los tres grupos de conjuntos de datos presentan diferentes niveles de datos: la primera categoría abarca información de convocatorias y cada convocatoria puede abarcar varias concesiones. La segunda categoría incluye el nivel de concesión y puede vincularse al conjunto de datos principal BDNS_CONCESIONES mediante ID de concesiones únicas. Por último, la última categoría es sobre

beneficiarios y el mismo beneficiario puede recibir varias concesiones. Por tanto, con el fin de fusionar todos los conjuntos de datos entre sí, el nivel de concesión se utilizó como unidad principal de análisis, proporcionando ID únicas.

La lista de variables relevantes para la evaluación de riesgos de fraude podría dividirse en indicadores de antecedentes e indicadores de riesgo. Se necesitan indicadores de antecedentes para describir las características específicas de las convocatorias, los concedentes, los beneficiarios y terceros que están potencialmente asociados a las sanciones. Los indicadores de riesgo se refieren a determinadas fases de publicidad, selección, ejecución y seguimiento de subvenciones. La Tabla 2.1 muestra la lista completa de indicadores de antecedentes, la Tabla 2.2 muestra los indicadores de riesgo que podrían extraerse de los conjuntos de datos de la IGAE.

Tabla 2.1. Indicadores de antecedentes

Grupo de indicadores	Nombre del indicador	Código de variables	Encabezado de variables
Concedente	Organizador de convocatorias	CON 705; CON 710; CSU 100	Entidad organizadora; Entidad concedente; Entidad concedente
Concesionario	Beneficiario	CON 580; CSU 120	Tipos de beneficiario; ID del beneficiario
	ID de la concesión	PRO 110; PAG 100; DEV 100; REI 100	Identificación de la concesión
	ID del proyecto	PRO 130; EJE 110	Identificación del proyecto
	Descripción del proyecto	PRO 210	Descripción del proyecto
	localización del proyecto	PRO 260	Región geográfica (proyecto)
	Id de la ejecución	EJE 120	Identificación del ejecutor
	año	EJE 130	Año de ejecución
	ID descalificada	INH 100	ID descalificada
	Fecha de inhabilitación	INH 210	Fecha de descalificación
	Entidad inhabilitadora	INH 220	Identificar el origen administrativo o judicial de la entidad incapacitante
	Período de inhabilitación	INH 230; INH 240	Fecha de inicio de la inhabilitación; Fecha de finalización de la inhabilitación
Subvención	valor de la subvención	CSU 220; EJE 210	Importe de la subvención; Importe de la subvención a la entidad ejecutora por año
	base reguladora de la subvención	CON 250; CON 260	Descripción BBBR; URL BBBR
	Identificación de la convocatoria	CON 290	Título de la convocatoria
	publicación de la convocatoria	CON 300; CON 310	Enviado para publicación; Fuente oficial
		CON 335	Título en español, Texto en español
		CON 335; CON 340	Fecha de publicación; Enlace a la publicación
	fecha de la firma y lugar	CON 351; CON 352	Fecha de la firma; Ubicación de la firma
	solicitud	CON 440; CON 460	Fecha de inicio; Fecha de finalización
	ayuda estatal	CON 490; CON 495; CON 515	Condición de la ayuda estatal; autorización de la ayuda; Identificador de la ayuda de la UE
	sectores de convocatorias	CON 550	Sectores de Economía
	ubicación de la convocatoria	CON 570	Regiones geográficas
	plazos	CON 600	Plazo para justificar la concesión
	subvención nominativa	CON 610	Concesión de carácter nominativo
	Fondos de la UE	CON 690	Importe de financiación del fondo de la UE
	normativa	CSU 110	ID de la normativa
	fecha de pago	PÁG 210	Fecha de pago
	importe pagado	PÁG 220	Importe de pago
	retención	PÁG 230	Retención de impuestos
	fecha de devolución	DEV 210	Fechad e devolución

Grupo de indicadores	Nombre del indicador	Código de variables	Encabezado de variables
	importe de devolución	DEV 220	Importe de devolución
	tipo de interés	DEV 230	Importe del tipo de interés
	fecha de reembolso	REI 210	Fecha de reembolso
	motivo de reembolso	REI 220	El motivo del reembolso
	importe reembolsado	REI 230	El importe del reembolso
Tercero	país	TER 100; TER 250	País del tercero; País del domicilio
	id	TER 110	ID del tercero
	nombre	TER 240	Nombre del tercero; Nombre comercial del tercero
	apellido	TER 210	Primer apellido del tercero; Segundo apellido del tercero
	dirección	TER 252; TER 254; TER 256; TER 258; TER 310	Dirección del domicilio; Código Postal; Municipio; Provincia; Región
	tipo	TER 280; TER 290	Condición jurídica; Tipo de tercero
	actividad	TER 320	Sector de Economía

Fuente: Autor

Tabla 2.2. Indicadores de riesgo

Fase 2	Nombre del indicador	Definición del indicador	Variables (código)	Variables (encabezado)
Competencia	Falta de publicidad	No hay publicidad electrónica adecuada del programa de subvenciones	CON 310; CON 420; CON 620	Fuente oficial; Solicitud abierta; Condiciones de publicidad de la concesión
Selección	Proceso de selección	Norma y proceso inadecuados para la selección	CON 560; CON 540; SAN 110; SAN 100; SAN 210	Herramienta de ayuda; Fin público; Discriminador de sanciones; Identificación del sancionado; Fecha de resolución de la sanción
	Selección inadecuada	Selección inadecuada de beneficiarios de subsidios	CSU 120; CSU 130; PAG 110; DEV 110; INH 110; CON 630	Beneficiario; Discriminador de concesión de subvenciones; Discriminador de pagos; Discriminador de devoluciones; Discriminador de descalificaciones; Impacto de género
Ejecución	Transacciones no supervisadas	Transacciones que eluden los procedimientos de revisión normales o que no se controlan de otro modo	CON 502; CON 503; CON 504	Exención del Reglamento de la UE por categoría de ayuda; Objetivos de la exención; Regulación de exención por importe
	Pagos inconsistentes	El pago es excesivamente caro o no está relacionado con los objetivos del programa de subvenciones.	CSU 250; CSU 240; CSU 220; PRO 220; PRO 240: PRO 250; EJE 220; EJE 240; EJE 250	Ayuda equivalente a la concesión de subvenciones; Coste financiable de la actividad (subvención); Importe de la subvención; Importe de la subvención para el proyecto; Costes del proyecto; Ayuda equivalente (proyecto); Importe de la subvención para el organismo ejecutor por año; Coste del proyecto asignado al organismo ejecutor por año; Ayuda equivalente (ejecutor)
	Pagos redondeados	Un beneficiario de una subvención de reembolso que extrae fondos de la subvención utilizando números redondeados a la centena, millar o superior más cercana puede indicar que los fondos no se están extrayendo sobre una base de reembolso.	CSU 250; CSU 240; CSU 220	Ayuda equivalente a la concesión de subvenciones; Coste financiable de la actividad (subvención); Importe de la subvención

Fase 2	Nombre del indicador	Definición del indicador	Variables (código)	Variables (encabezado)
Seguimiento	Sanciones	Número elevado de infracciones sistemáticas por parte del destinatario	SAN 250, SAN 280; SAN 440; SAN 450	Multa por infracciones menores; Multa por infracciones graves; Publicidad de sanción; Plazo para publicitar la sanción

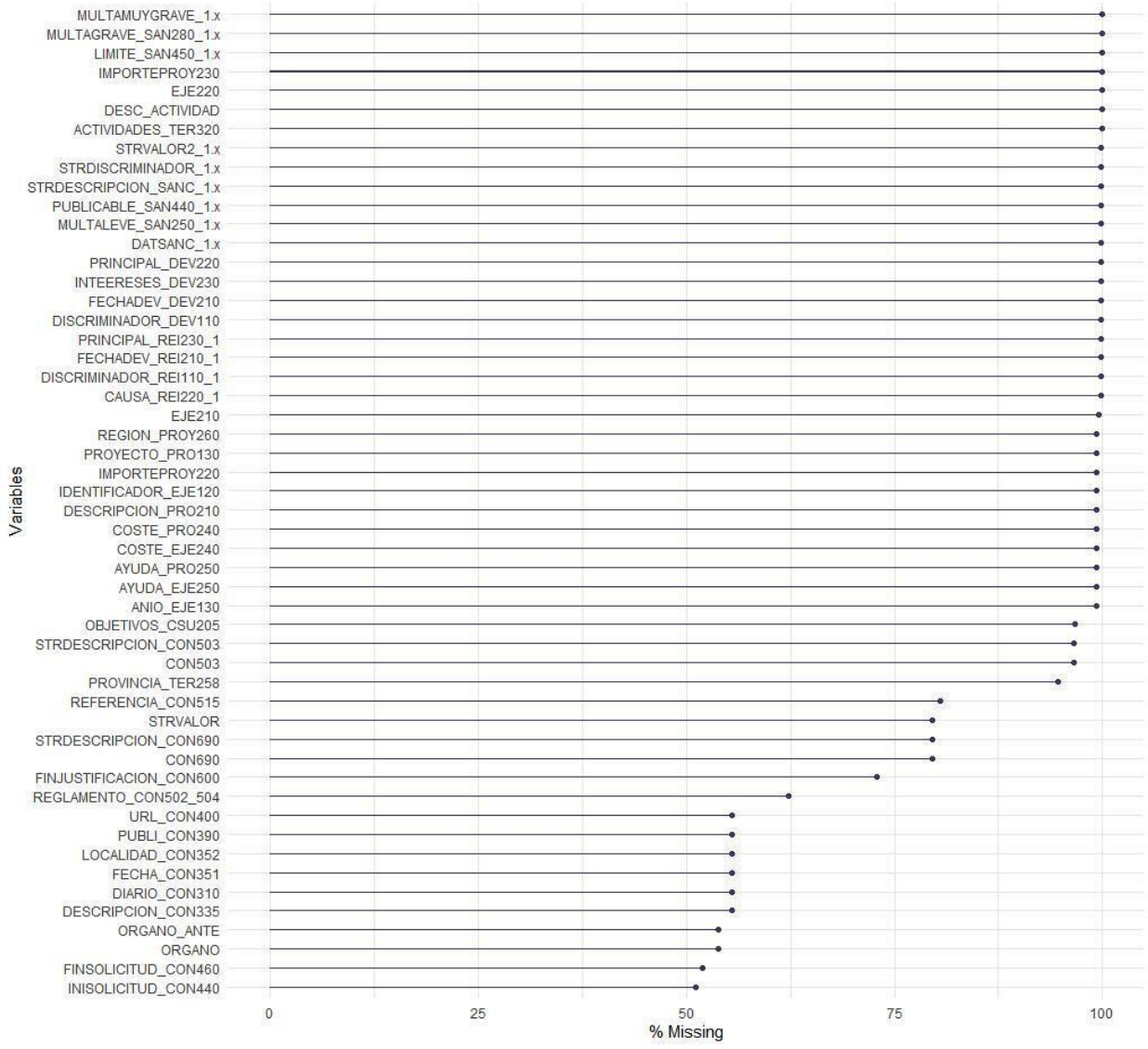
Fuente: Autor

Fusionar, limpiar y comprender las limitaciones de los datos

El primer paso del tratamiento de datos ha sido fusionar todos los conjuntos de datos facilitados por la administración española en el conjunto de datos principal que abarca todas las convocatorias y concesiones, BDNS_CONCESIONES. Para hacer eso, cada conjunto de datos se alineó con la misma unidad de análisis: la ID de la concesión. Cuando varias observaciones estaban relacionadas con la misma ID de concesión (por ejemplo, una concesión relacionada con varios sectores económicos), los datos se agregaban, por ejemplo, colocando cada observación duplicada en una columna aparte. Cuando las observaciones estaban relacionadas con una serie de ID de concesiones (por ejemplo, cuando el conjunto de datos contenía información sobre convocatorias), las características relevantes se copiaban en todas las concesiones relacionadas con esa observación de nivel superior. El conjunto de datos combinado, pero sin limpiar, contiene 1 792 546 concesiones y 152 variables. La lista completa de variables incluidas se puede encontrar en el Anexo B.

El siguiente paso del tratamiento de datos fue la limpieza de datos. Esto implicó eliminar variables con un índice alto de omisión o baja varianza. Estos problemas de datos afectan a un gran número de variables, como se muestra en la Figura 2.1. Se eliminaron todas las variables con un índice de omisión superior al 50 %, ya que habría introducido un mucho ruido en el análisis. La mayoría de estas variables con un índice alto de valores omitidos corresponden a sanciones y descripción del proyecto. Además, algunas de las variables mostraron una varianza muy baja, inferior a 0,3, lo que significa que contienen muy poca información relevante para el análisis posterior (es decir, en términos técnicos: su valor discriminante es bajo, ya que no varían lo suficiente entre observaciones sancionadas y no sancionadas). Por último, se eliminaron las variables de texto que no son directamente relevantes para la calificación de riesgo, como los descriptores de texto de las variables categóricas (por ejemplo, descripciones de sector económico) y las variables de texto libre con poca información relevante (por ejemplo, el título de la convocatoria).

Figura 2.1. Índices de valores omitidos



Fuente: Autor.

Como los métodos analíticos utilizados pueden ser sensibles a la información omitida, solo se conservaron aquellas observaciones que no tenían valores omitidos en todas las variables consideradas en el análisis, como las concesiones. Después de realizar todos estos pasos de tratamiento de datos, el conjunto de datos final utilizado en el análisis consta de 1 050 470 observaciones, concesiones y 60 variables para el periodo de 2018 a 2020 (inclusive).

Usar datos existentes para crear nuevos indicadores

Si bien la mayoría de los indicadores utilizados en el análisis se derivan directamente de los datos recibidos, algunos indicadores también se calcularon combinando otras variables. El primer grupo de estos indicadores calculados se refiere al importe y número de concesiones recibidas por el mismo beneficiario. El segundo grupo está formado por variables relacionadas con la ubicación: el nivel territorial del concedente y el beneficiario: nacional, regional o local. Además, se creó una variable binaria para identificar si la ejecución del proyecto se ubicó en el mismo lugar que el tercero. En tercer lugar, se calculó

un indicador que captura el mes de concesión de la subvención que puede indicar la periodicidad del gasto y los riesgos correspondientes. Por último, el sector económico del beneficiario se agrupó para recoger solo el nivel más alto de la clasificación NACE (Sección, categorías de 1 dígito). Véase en el Anexo A una tabla que describe todas las variables de estos cálculos de indicadores adicionales, y los pasos de tratamiento de datos detallados anteriormente. Esta es la lista final de variables utilizadas para el modelado de riesgos.

Definición de la variable dependiente en función del estado sancionador

La principal variable dependiente utilizada para el análisis es una variable binaria que indica si el beneficiario que recibe la concesión ha sido sancionado o no; con la sanción interpretada como una indicación fiable de fraude en una concesión. La variable pasa al valor «1» si el beneficiario ha sido sancionado por la concesión correspondiente, así como por todas las concesiones anteriores recibidas por el mismo, por haberse producido prácticas fraudulentas con anterioridad a la fecha de sanción. En caso de que el tercero no fuera sancionado, la concesión correspondiente obtiene el valor «0» en la variable ficticia. Las clases en la variable de sanción están muy desequilibradas: muestra 1031 casos de sanciones frente a 1 049 439 casos de ausencia de sanciones.

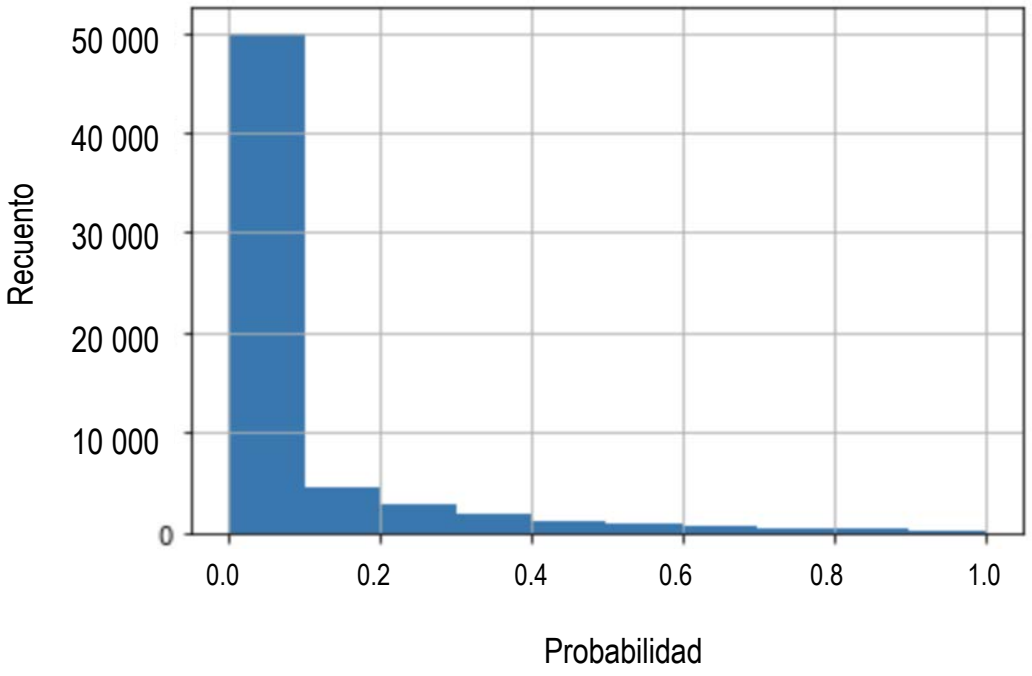
Para que el algoritmo de *random forest* se ejecute de manera eficiente, se ha extraído una muestra aleatoria de 90 000 concesiones de la parte sin etiquetar del conjunto de datos. Por tanto, el conjunto de datos de entrenamiento utilizado en el análisis hace uso de la muestra inicial de 91 031 concesiones, que consta de 1031 concesiones positivas (sanciones conocidas) y 90 000 concesiones sin etiquetar (estado de fraude poco claro).

Asignar el estado de sanción a las concesiones sin etiquetar

Para asignar etiquetas positivas y negativas a las observaciones sin etiquetar, se utilizó la metodología de aprendizaje positivo sin etiquetar. Este método comienza creando un subconjunto de entrenamiento de los datos que consta de todos los casos positivos, y una muestra aleatoria de casos sin etiquetar. Sobre esta muestra, el *bagging* PU construye un clasificador que asigna la probabilidad de sanción a cada concesión, a partir del cual es posible asignar la etiqueta positiva y negativa (probabilidad de sanción >50 % → etiqueta positiva). Estos pasos se repiten 1000 veces para construir un clasificador fiable que identifique los casos probables negativos y probables positivos en la muestra sin etiquetar (se debe tener en cuenta que la probabilidad de sanción media pronosticada en todos los modelos se convertirá en la calificación final pronosticada).

Como resultado de la ejecución de estos algoritmos, todos los casos sin etiquetar han recibido una probabilidad de sanción y, por tanto, una etiqueta de sanción probable (positiva frente a negativa). Para el conjunto de datos de entrenamiento, la Figura 2.2 presenta la distribución de las probabilidades de sanción (es decir, fraude). Esto pone de manifiesto que la mayoría de concesiones se consideran de bajo a muy bajo riesgo, y solo unas cuantas reciben una calificación de alto riesgo. En otras palabras, la mayoría de concesiones pueden clasificarse como no sancionadas, mientras que muy pocas concesiones reciben la etiqueta de sancionadas. En comparación con la muestra inicial positiva sin etiquetar, el número de casos probables positivos (sancionados) aumentó a 4430 con 86 601 identificados como probablemente negativos (no sancionados).

Figura 2.2. Clasificador de *insaculación* PU: predicción de probabilidad de sanción en la muestra inicial



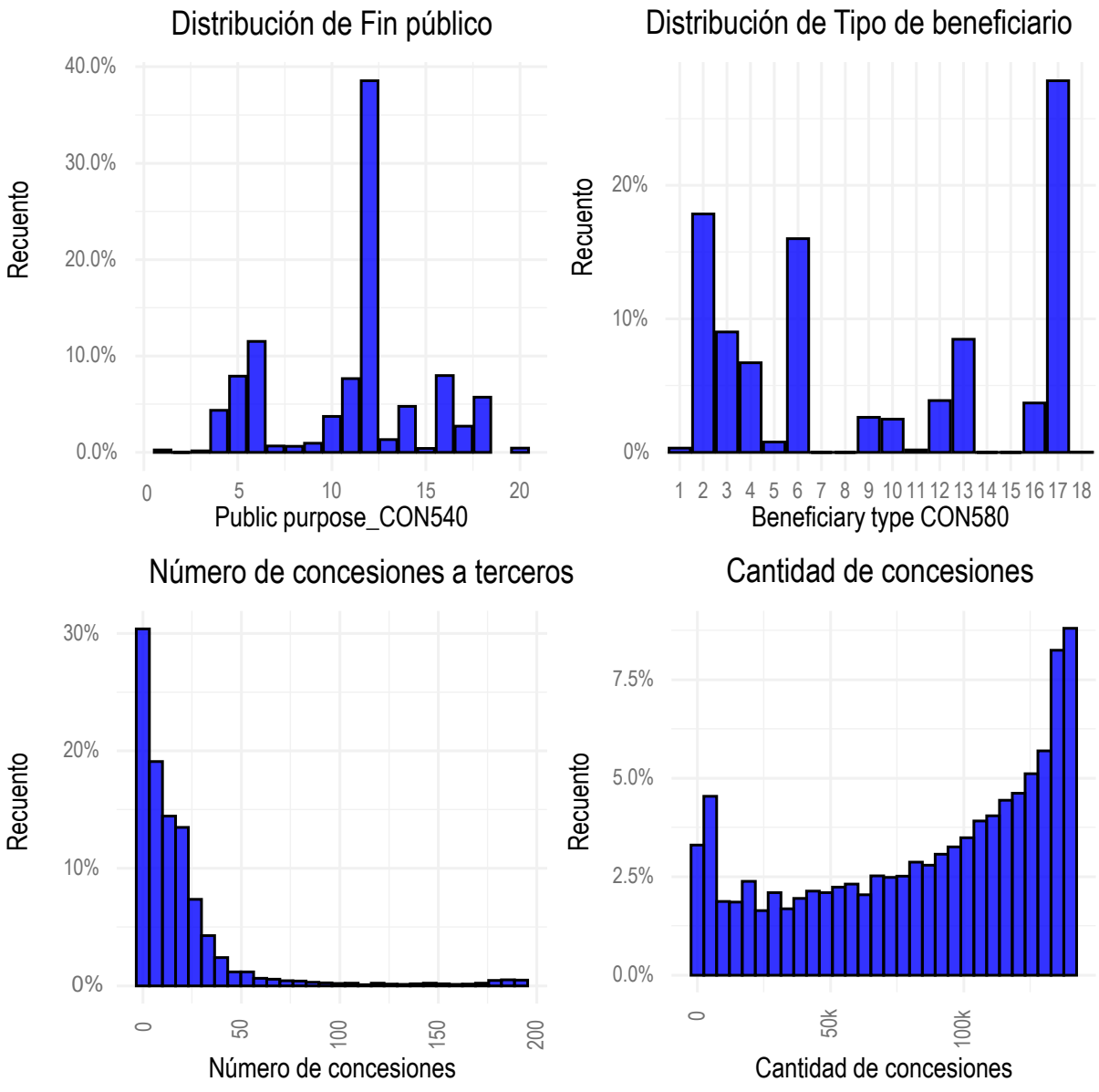
Fuente: Autor

Identificar las variables de mayor más impacto

Una vez que el conjunto de datos de concesiones positivos y sin etiquetar se vuelve a etiquetar y solo quedan casos positivos y negativos en el conjunto de datos siguiendo los métodos anteriores, se ejecuta un nuevo modelo de *random forest* y se comprueba su precisión. Esto significa que el conjunto de datos reetiquetado de 91 031 concesiones se dividió en una muestra de entrenamiento (70 %) y una muestra de verificación (30 %). El algoritmo *Random Forest* se entrenará en la primera y probará su precisión en la otra muestra que no ha «visto». El modelo óptimo consta de 1000 árboles y utiliza 106 variables en cada ejecución.

Este modelo óptimo de *random forest* ha identificado las variables más importantes para predecir la probabilidad de sanciones. A efectos de modelización, cada variable categórica se transformó en un conjunto de variables binarias, de modo que correspondan a una sola categoría de la variable categórica. Las variables numéricas se utilizaron tal cual, sin transformación. Las variables de mayor impacto en el modelo óptimo de *Random Forest* son Public_purpose_CON540, Nawards_TER_110, Amount_awards_TER110 y Third_party_legal_Spain_TER280. Sus distribuciones se presentan en la Figura 2.3.

Figura 2.3. Distribuciones de las variables de mayor impacto



Fuente: Autor

Estas distribuciones muestran que muchas de las variables más importantes tienen distribuciones significativamente desiguales. Por ejemplo, el número de subvenciones cae estrepitosamente por debajo de 50 y muy pocos beneficiarios tienen más de 50 subvenciones concedidas. De manera similar, la variable de fin público tiene un pequeño número de categorías predominantes, como 12 (agricultura). Además, el número de concesiones recibidas por el mismo beneficiario no está correlacionado con el valor total de las concesiones, lo que significa que la cantidad media de concesiones distribuidas es relativamente baja y algunas concesiones tienen un valor muy alto. La siguiente sección va un paso más allá y analiza los impactos de estas variables en la probabilidad de sanción (fraude).

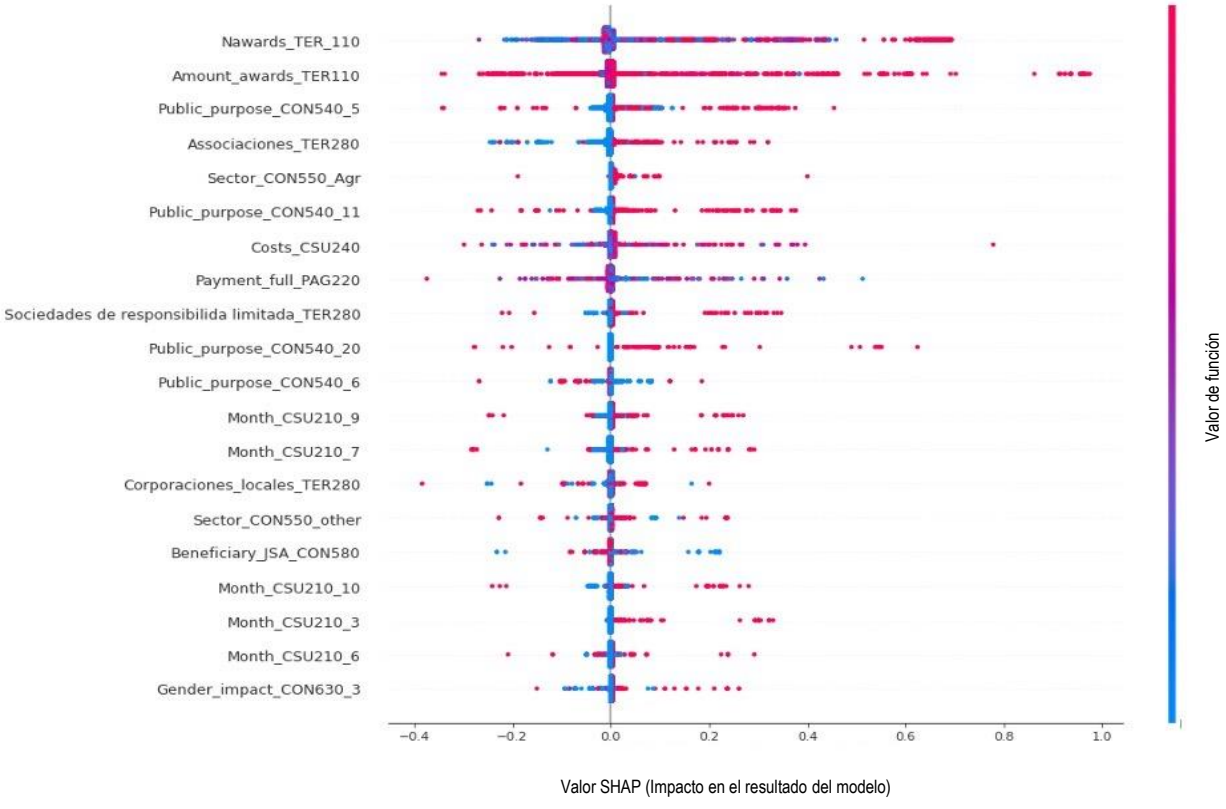
Probar el modelo en un conjunto de datos ciego

El modelo óptimo entrenado en el conjunto de datos de entrenamiento se ha verificado sobre datos no usados anteriormente, el conjunto de prueba (30 % de la muestra). Sobre este conjunto de datos de prueba, el modelo óptimo de *random forest* alcanzó: [5]

- exactitud = 95 % (la exactitud es el número de etiquetas predichas correctamente de todas las predicciones realizadas), y
- repetición = 93 % (es el número de etiquetas que el clasificador identificó correctamente dividido por el número total de observaciones con la misma etiqueta).

Estos resultados nos llevan a la conclusión de que el modelo es de gran calidad. Después de determinar la calidad general del modelo, la atención se ha centrado en el impacto de los predictores individuales en la probabilidad de sanción (fraude). Se debe tener en cuenta que los modelos de *Random Forest* capturan una gama de efectos interactivos y no lineales, por lo que interpretar las relaciones entre los predictores y el resultado es un asunto polifacético y complejo. Para mostrar el impacto de cada predictor de impacto en el resultado del modelo se siguió la literatura más reciente sobre aprendizaje automático, se calcularon los valores de explicaciones aditivas de Shapley (SHAP) (Lundberg and Lee, 2017[6]) y se han representado gráficamente. Los valores de SHAP ayudan a identificar la contribución individual de cada característica al modelo y su importancia para la predicción. El gráfico de Shapley en la Figura 2.4 muestra la probabilidad de sanciones (es decir, probable fraude) en función de los diferentes valores de cada predictor de impacto.

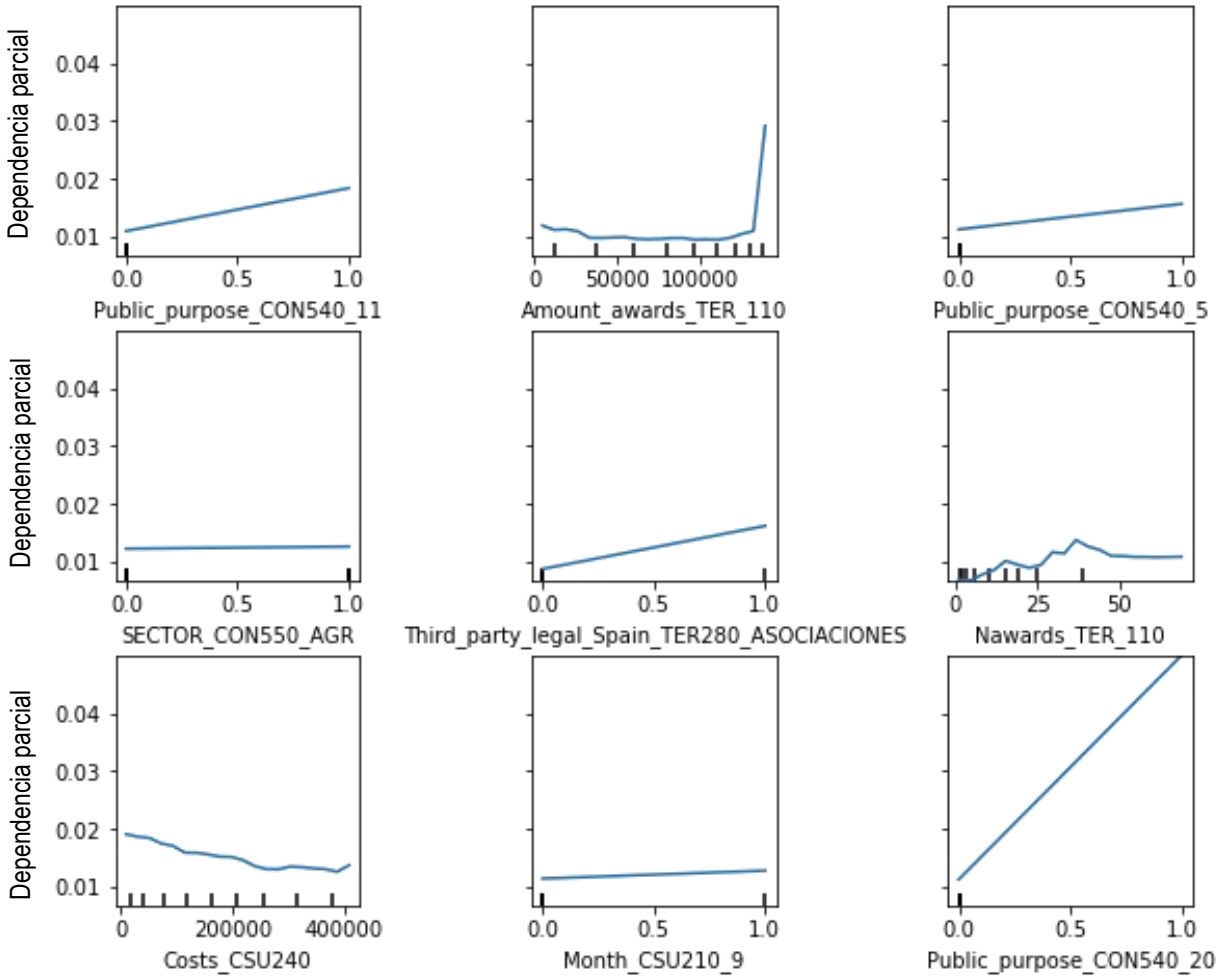
Figura 2.4. Valores de SHAP: Importancia variable y dirección del efecto



Fuente: Autor

La Figura 2.4 destaca que el impacto positivo más significativo en la probabilidad de sanciones lo proporciona el número de concesiones, así como el valor total de las concesiones recibida por el mismo beneficiario. Respecto al resto de predictores, la probabilidad de sanciones se correlaciona positivamente con la asociación y las sociedades limitadas como forma jurídica de beneficiarios, así como con el sector agrario en el sector económico. Los importes de la subvención están asociados negativamente a la probabilidad de ser sancionados, lo que significa que los importes más altos de los proyectos no se correlacionan con riesgos más elevados. Por el contrario, los fines públicos de la concesión, como la cultura (11), los servicios sociales (5), la cooperación internacional para el desarrollo y la cultura (20) y el fomento de empleo (6) están relacionados con una mayor probabilidad de sanciones. Además, las subvenciones concedidas en septiembre y julio se asocian a mayor probabilidad de sanción, con una tendencia similar en octubre, marzo y junio. Se muestran visualizaciones más detalladas de la influencia de las variables importantes en la probabilidad de sanciones (fraude) en la Figura 2.5.

Figura 2.5. Gráficos de dependencia parcial que representan el impacto de las variables seleccionadas en la probabilidad de fraude



Fuente: Autor

Finalización de la lista de indicadores para el modelo de riesgo

Para completar la descripción del modelo de evaluación de riesgos, se incluye el listado final de 29 indicadores válidos utilizados por el modelo según seis grupos (Tabla 2.3), haciendo referencia a las fases en las que podría ocurrir el posible fraude o las características de las organizaciones participantes: Fases de competición, selección, ejecución y seguimiento; organismo que concede la subvención y organización destinataria (beneficiario).

Tabla 2.3. Lista final de indicadores

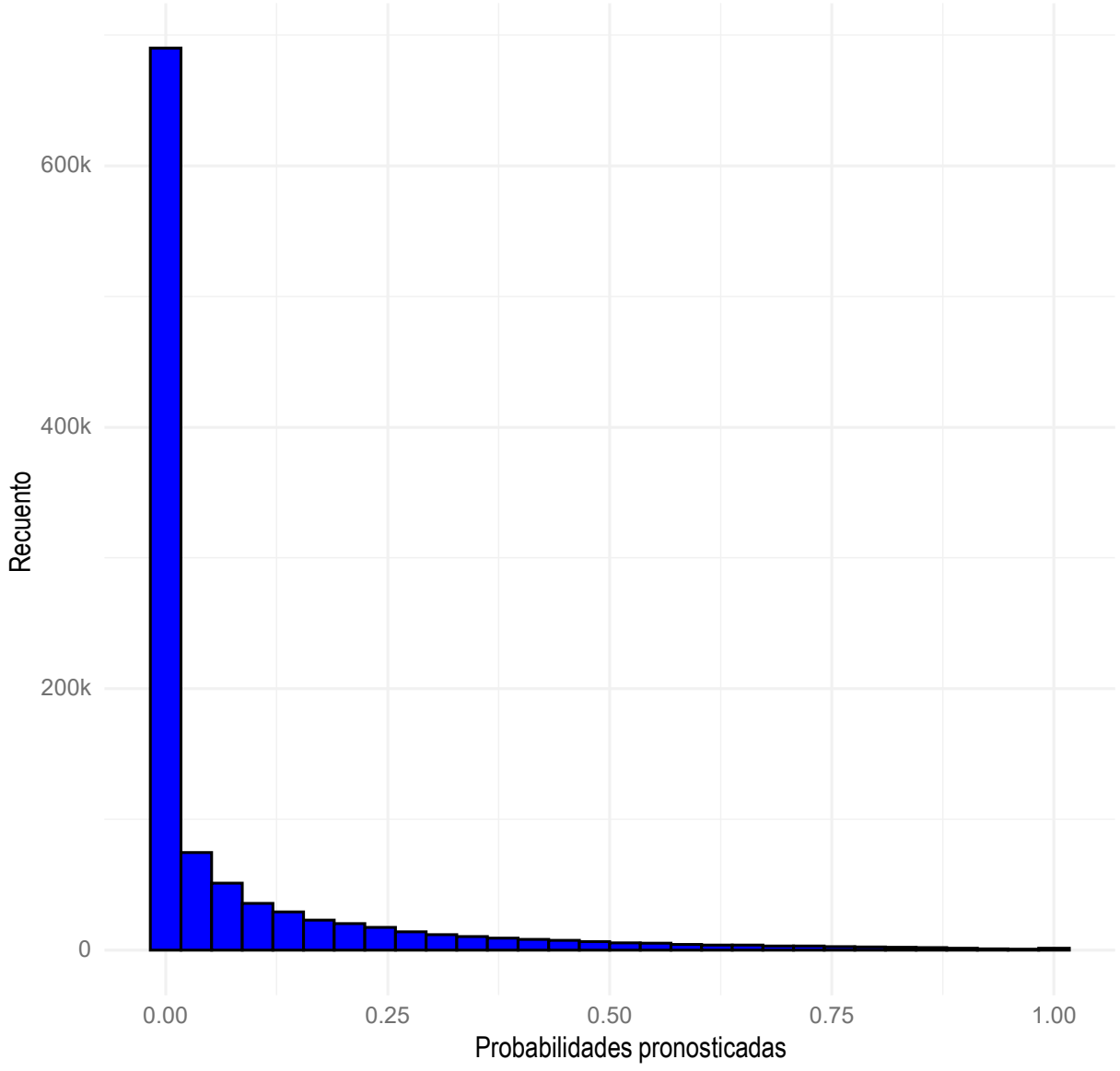
Fase	Variable	Descripción de la variable	Fraudes
Fase de competición	CON420, CON490, CON620	Admisión abierta, Condición de ayuda estatal, Convocatoria pública	La ausencia de admisión abierta o convocatoria pública conduce a un proceso de seguimiento menos transparente y, por tanto, existe mayor predisposición a actividades fraudulentas.
Fase de selección	CON540, CON580, CON610, CON630, SECTOR_CON550_AGR...EXTRATER, Month_CSU210	Fin público, Tipo de beneficiario, Subvención nominativa, Impacto de género, Sector de la economía, Mes de concesión	El tipo y la fecha de la convocatoria, el sector de la economía y el tipo de beneficiario podrían estar correlacionados con determinadas prácticas fraudulentas
Subvención/ejecución	CSU240, CSU220, CSU250, PAG220, PAG230, CON560, LOCAL_IMPL	Subvención nominativa, Importes, Concesión de subvención, Ayuda, Importe total pagado, Retención de impuestos, Instrumento de ayuda, Implantación local	Las subvenciones de importe elevado podrían ser potencialmente más propensas a actividades fraudulentas. Si la implantación se lleva a cabo en el mismo lugar que el concedente, podría ser una señal de un esquema de corrupción.
Organismo concedente	NATIONAL_CSU260, REGIONAL_CSU260, MUNICIAPAL_CSU260	Nivel de concesión de subvenciones	Las capacidades administrativas en determinadas regiones podrían ser insuficientes para un seguimiento eficaz de la convocatoria.
Organización destinataria	TER100, TER250, TER280, TER290, NATIONAL_TER310, REGIONAL_TER310, MUNICIAPAL_TER310, Amount_awards_TER110, Naward_TER110	País de terceros, Ubicación de terceros, Naturaleza jurídica de terceros, Tipo de terceros, Nivel de terceros, Número de concesiones, Cantidad de concesiones	La estructura y el tipo de organización de terceros, así como la ubicación, podrían estar correlacionadas con actividades fraudulentas. Las partes que reciben más concesiones de mayor tamaño podrían ser potencialmente más fraudulentas que otras.
Seguimiento	SAN_dum	Concesiones sancionadas	Captura la actividad fraudulenta del tercero

Fuente: Autor

Presentación de resultados y consideraciones para un desarrollo ulterior

La potencia de la metodología de evaluación de riesgos propuesta se muestra mejor utilizando el modelo óptimo final de *Random Forest* para asignar una calificación de riesgo de fraude a todas las concesiones de 2018 a 2020 con suficiente calidad de datos. Por tanto, la distribución final de la probabilidad de sanciones se presenta en la Figura 2.6 para las 1 050 470 concesiones observadas. En esta amplia muestra, el modelo predice que no habrá fraude (sanciones = 0) para 1 008 318 concesiones, mientras que predice fraude (sanciones = 1) para 42 152 concesiones utilizando el umbral del 50 % de probabilidad de sanciones para distinguir entre sanciones y no sanciones.

Figura 2.6. Distribución de probabilidades pronosticadas para todas las concesiones, nivel de concesión, 2018-2020

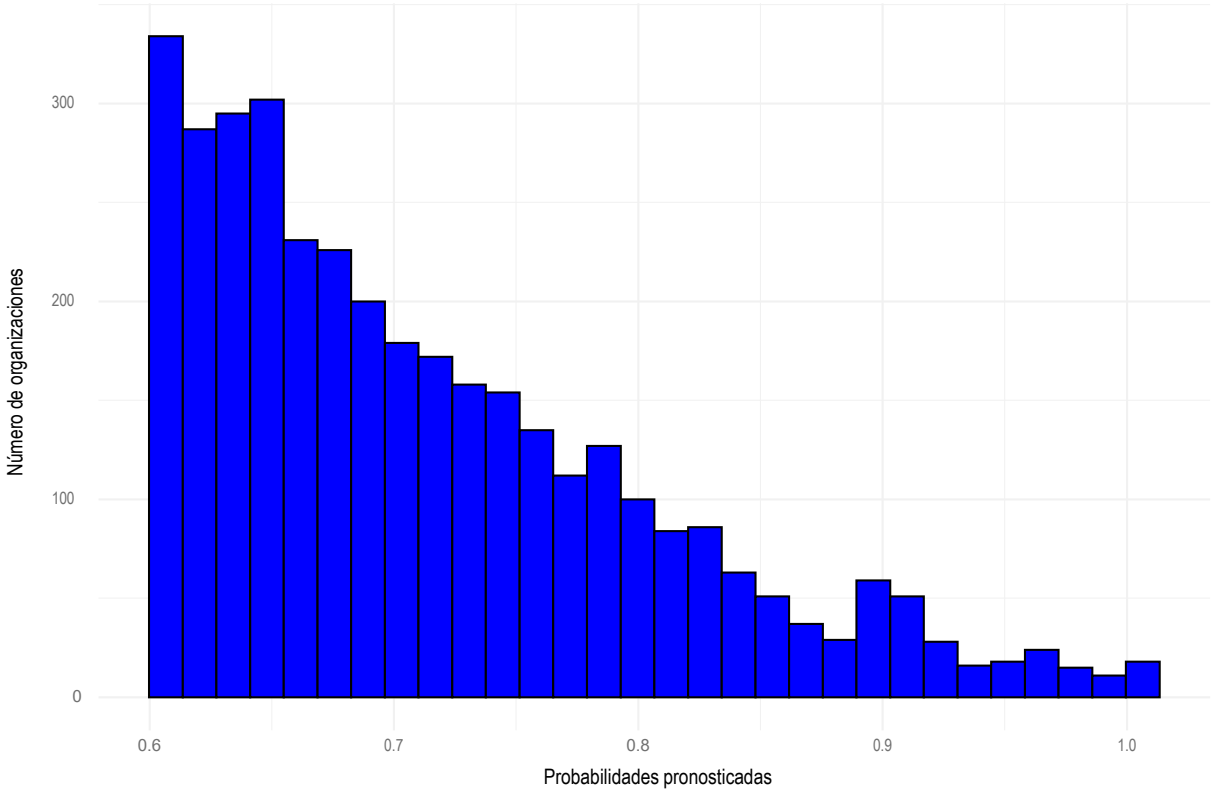


Fuente: Autor

Dado que los riesgos tienden a agruparse a nivel de las organizaciones y que las investigaciones suelen examinar todas las subvenciones recibidas por una organización, la exploración de las predicciones de las probabilidades de fraude a nivel de los beneficiarios añade valor al modelo. Para ofrecer una visión general de este nivel de agregación, mostramos la distribución de las predicciones de los riesgos de fraude por beneficiario con probabilidades de alto riesgo en la Figura 2.7. Ésta muestra que entre los beneficiarios de alto riesgo, las probabilidades de riesgo se distribuyen de forma desigual. La mayor parte de los beneficiarios de subvenciones de alto riesgo tienen características que indican una probabilidad de entre el 60% y el 70% de ser fraudulentos, y un grupo muy pequeño de organizaciones situadas en la cola derecha de la distribución tienen una probabilidad de ser fraudulentos de casi el 100% según el modelo. Estas organizaciones, las 10 primeras de las cuales se muestran en la Tabla 2.4, presentan el mayor riesgo y son las candidatas más adecuadas para un examen más profundo y una posible investigación

basada en el modelo predictivo. Además de éstas, las organizaciones a las que la IGAE decida seguir investigando dependerán de dónde fije su umbral de riesgo, y potencialmente de otros factores, como las implicaciones financieras (véase la sección « Combinar las puntuaciones de riesgo pronosticadas con la información financiera »)

Figura 2.7. Distribución media de probabilidades pronosticadas para organizaciones de alto riesgo, nivel de terceros, 2018-2020



Fuente: Autor

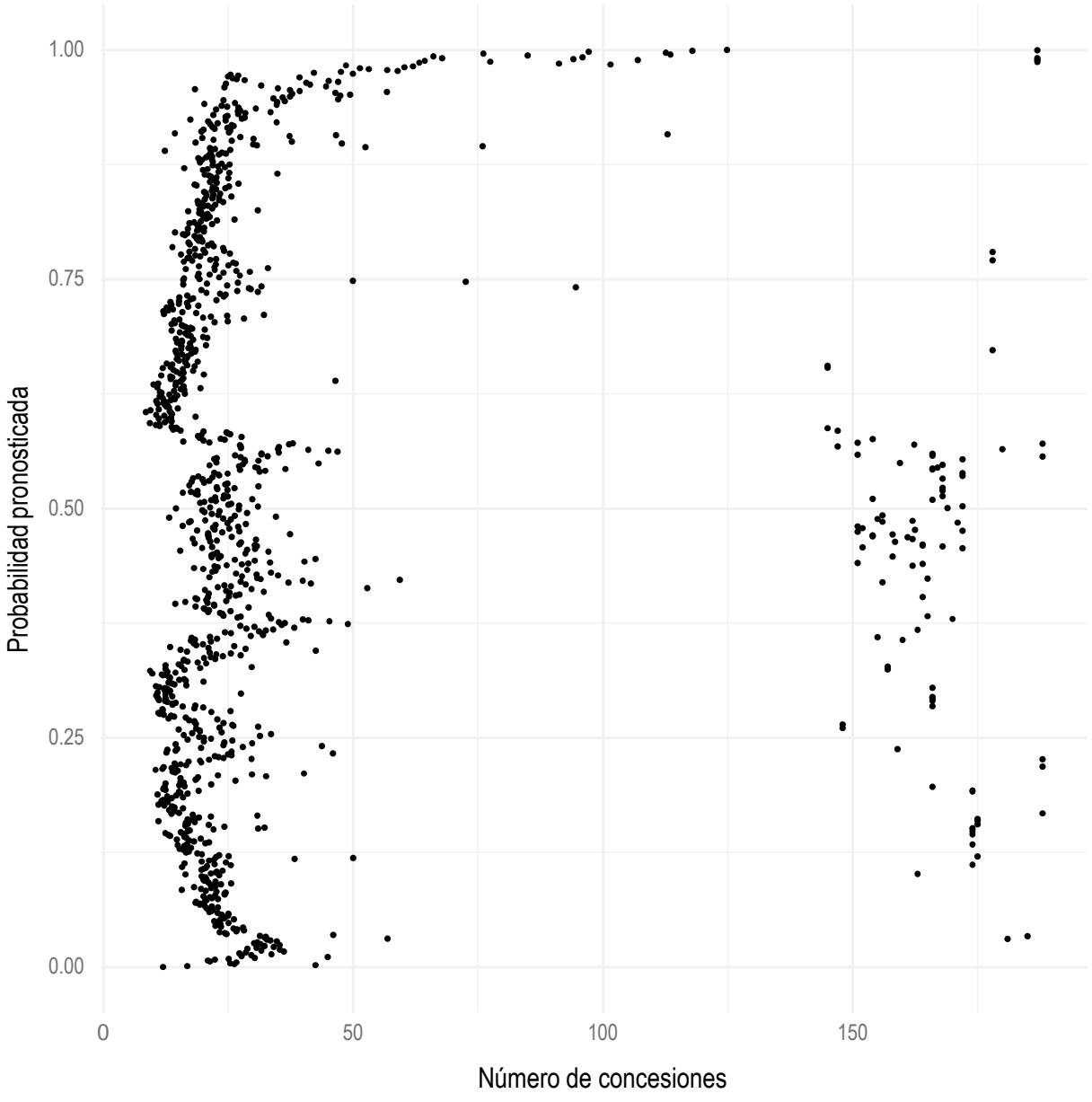
Tabla 2.4. Las 10 organizaciones principales por valor medio de concesiones

ID generada	Probabilidad pronosticada (media por tercero)
22568	1
46462	1
60626	1
101336	1
102140	1
129947	1
144235	0,996
152526	0,988
159661	1
167691	1

Fuente: Autor

Como era de esperar, el modelo estima que la inmensa mayoría de las concesiones no presentan riesgos, pero algunos miles de concesiones se marcan como de riesgo, además de las 1031 concesiones observadas sancionadas. Teniendo en cuenta las variables más importantes del modelo, se analiza más de cerca la distribución de las probabilidades de fraude. Primero, la Figura 2.8 muestra la distribución del número de concesiones recibidas por un mismo beneficiario en relación con su probabilidad de sanción. Curiosamente, el modelo predice una alta probabilidad de sanción para entidades receptoras grandes y pequeñas. La mayoría de concesiones se ubican en el lado izquierdo del gráfico, con 0 a 50 concesiones por beneficiarios y probabilidades relativamente igualadas de ser sancionadas para este grupo de observaciones. A partir de 50 concesiones, la probabilidad aumenta a casi el 100 %, con una disminución a alrededor del 50 % cuando el número supera las 150 concesiones. Esto podría explicarse por la fiabilidad de beneficiarios: si se demuestra que estas organizaciones son fiables durante largo tiempo reciben más concesiones. Mientras que para las primeras 50 concesiones, se lleva a cabo un proceso de evaluación. También es concebible que después de un umbral determinado de 50 concesiones por beneficiario, las investigaciones se lleven a cabo con mayor frecuencia y, por tanto, es más probable que aparezcan concesiones potencialmente sancionadas.

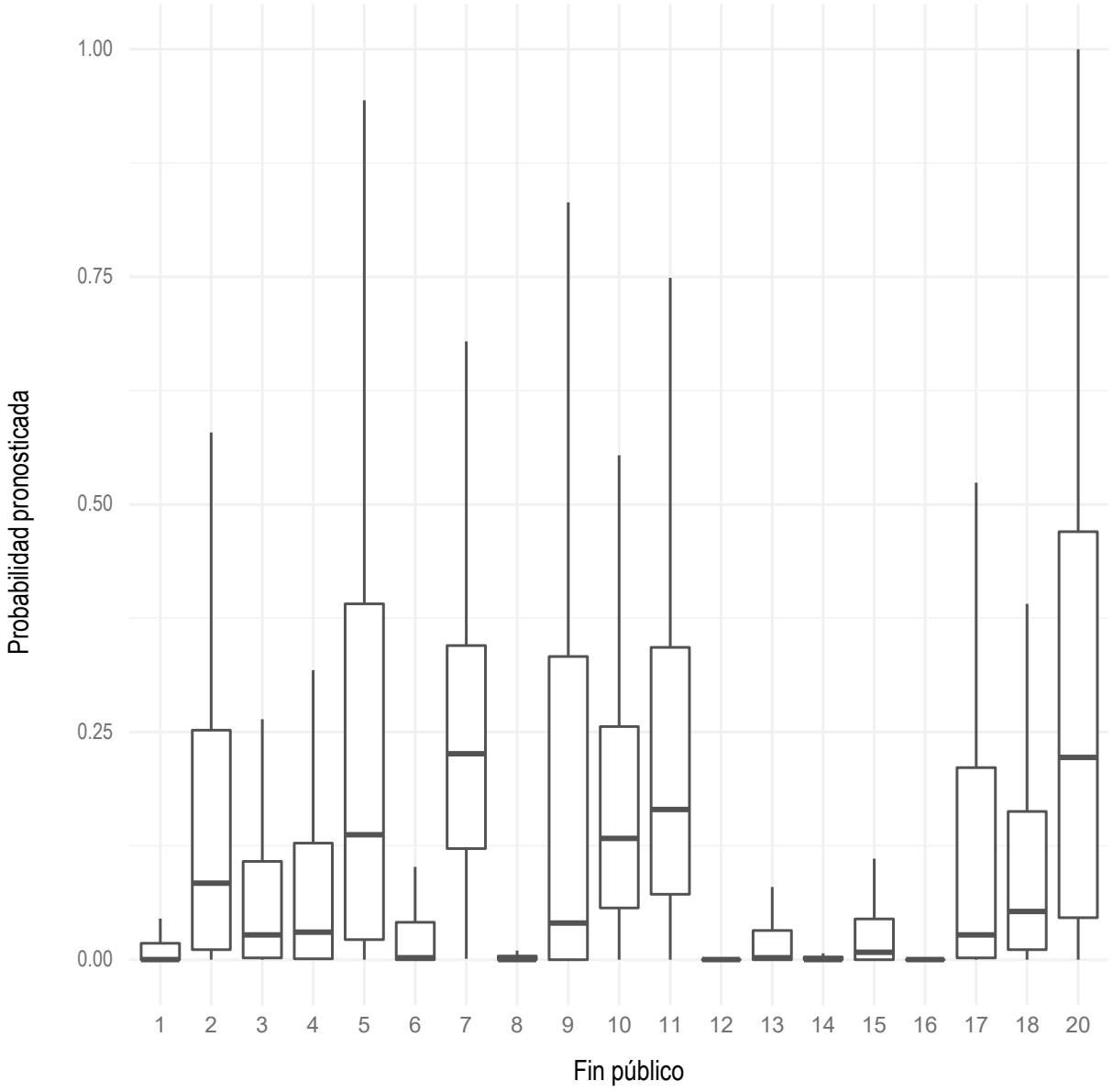
Figura 2.8. Distribución del número de concesiones por probabilidad de sanciones



Fuente: Autor

En segundo lugar, respecto a otra variable importante, el fin público de la convocatoria, se presenta su distribución de probabilidades en la Figura 2.9. Dos categorías muestran el mayor riesgo de sanciones: servicios sociales (5) y cooperación internacional para el desarrollo y la cultura (20). Es importante destacar que estas no son las categorías más frecuentes entre las concesiones; la más frecuente es la agricultura (12), que muestra el riesgo más bajo.

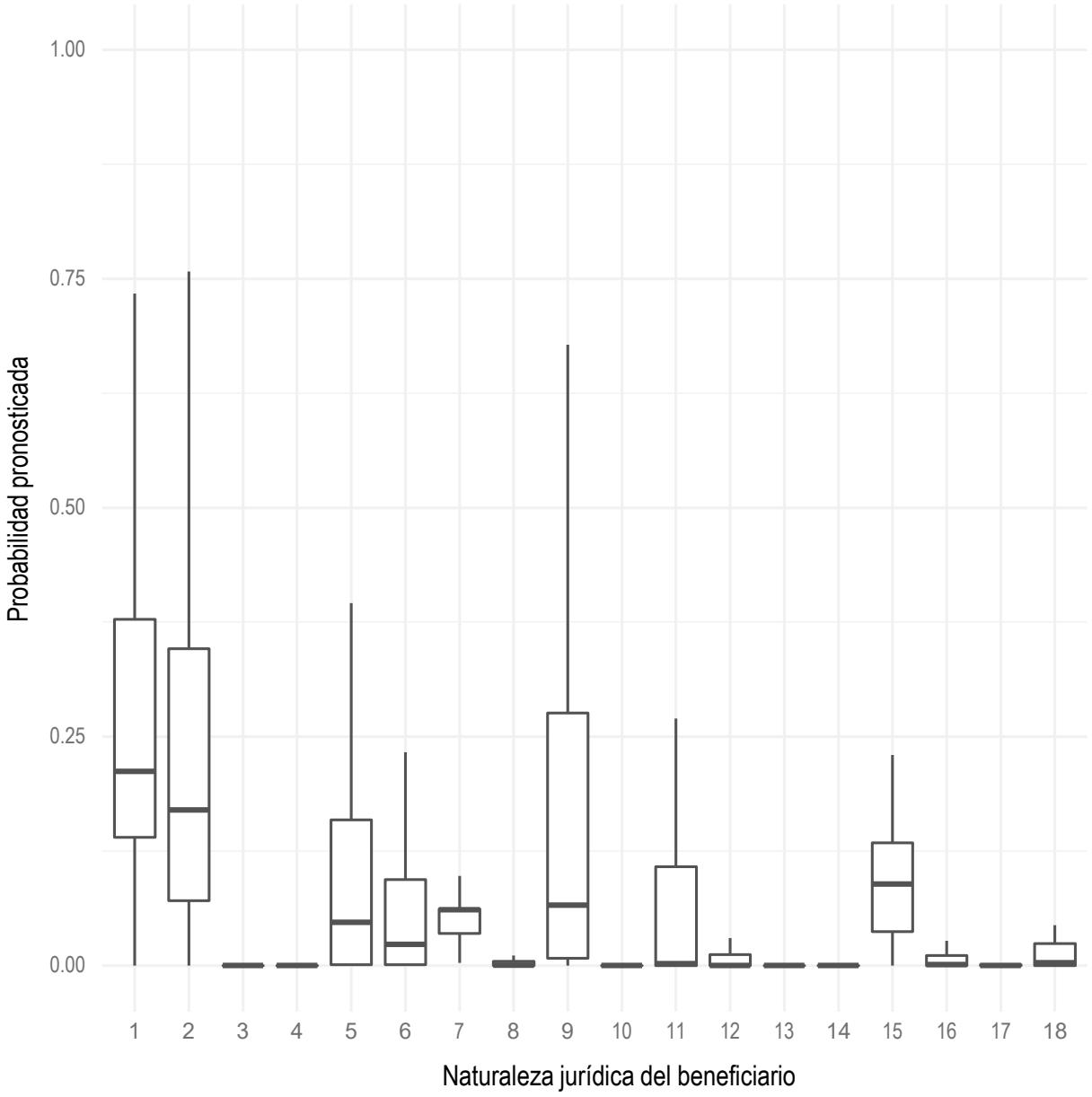
Figura 2.9. Distribución del fin público de la convocatoria sobre la probabilidad de sanciones



Nota: 1 - Justicia, 2- Defensa, 3 - Seguridad ciudadana e instituciones penitenciarias, 4- Otros beneficios económicos, 5- Servicios sociales y promoción social, 6- Fomento de empleo, 7- Desempleo, 8- Acceso a la vivienda, 9- Salud, 10- Educación, 11- Cultura, 12- Agricultura, Pesca y Alimentación, 13- Industria y Energía, 14- Comercio, Turismo y Pymes, 15- Subsidios para transporte, 16- Infraestructura, 17- Investigación, Desarrollo e Innovación, 18- Otras acciones económicas, 20 - Cooperación internacional para el desarrollo y la cultura
Fuente: Autor

En tercer lugar, la naturaleza jurídica del beneficiario es otra variable importante identificada por el modelo (Figura 2.10). La segunda categoría - asociaciones - mostró un impacto positivo significativo en la probabilidad de sanciones en el modelo presentado. Otros dos tipos de beneficiarios también son propensos a mayores riesgos: los órganos de la administración estatal y las comunidades autónomas (1) y los organismos públicos (9). Si bien la asociación es también la categoría más frecuente para esta variable, las categorías 1 y 9 son las menos frecuentes, pero muestran una gran probabilidad de ser sancionadas.

Figura 2.10. Distribución de la naturaleza jurídica de terceros sobre la probabilidad de sanciones

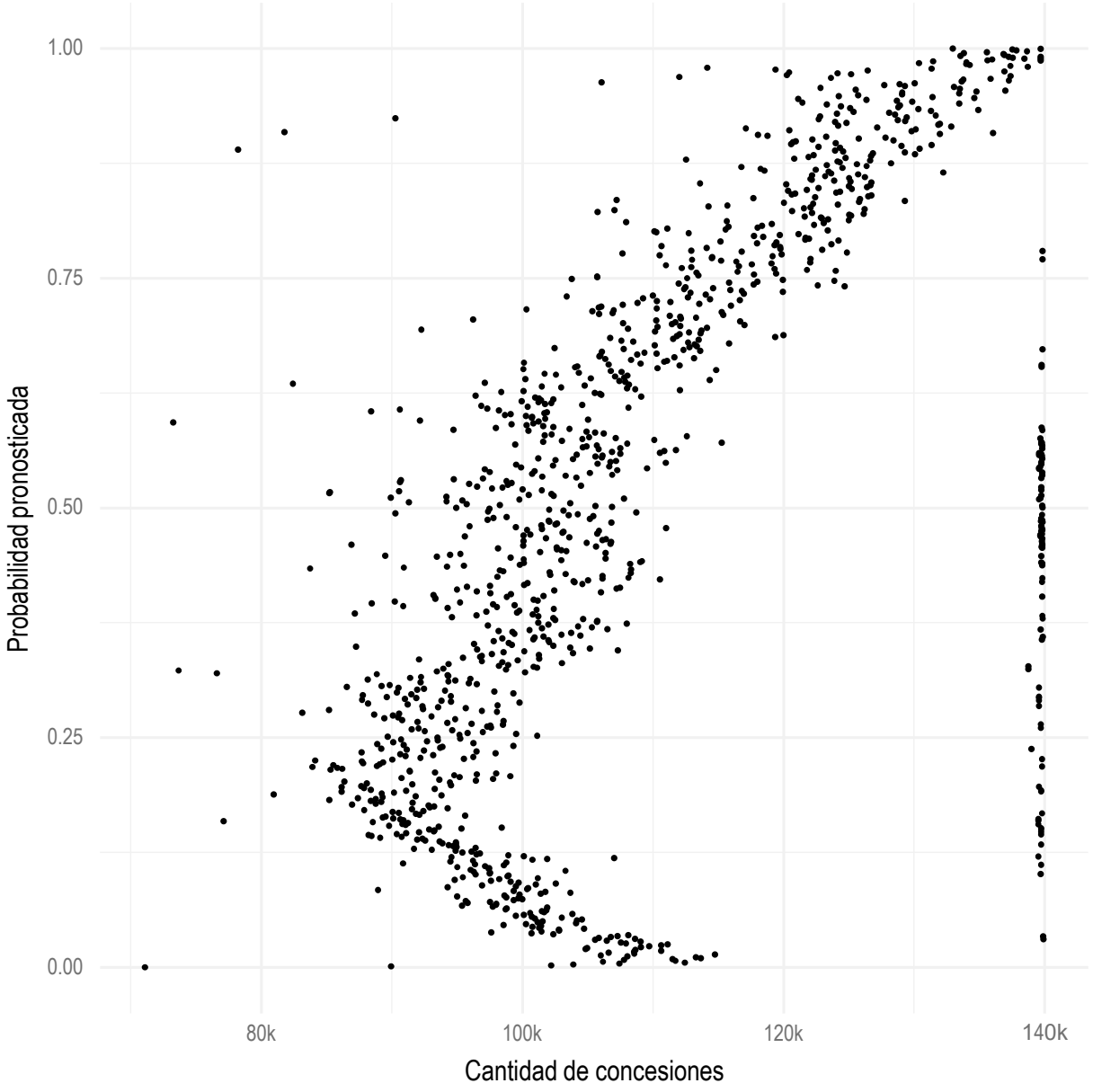


Nota: 1 - Órganos de la administración estatal y comunidades autónomas, 2 - Asociaciones, 3 - Comunidades de propiedad, herencias y otras entidades sin personalidad jurídica, 4 - Comunidades de propietarios en régimen de propiedad horizontal, 5 - Instituciones religiosas, 6 - Administración local, 7 - Entidad extranjera, 8 - Establecimiento permanente de entidad no residente en territorio español, 9 - Organismos públicos, 10 - Otros tipos, 11 - Persona jurídica con identificación no generada por autoridades españolas (AEAT o Policía), 12 - Sociedades anónimas, 13 - Organizaciones civiles, 14 - Organizaciones colectivas, 15 - Sociedades comandadas, 16 - Sociedades cooperativas, 17 - Sociedades limitadas, 18 - Uniones temporales de empresas
Fuente: Autor

Por último, también se ha encontrado que el importe total de las concesiones recibidas por el beneficiario tiene un impacto significativo en la probabilidad de sanciones (Figura 2.11). Hay un crecimiento sostenido de las probabilidades de sanción a partir de 90 000 €. Además, existe una divergencia en las probabilidades previstas entre 85 000 € y 110 000 €, lo que demuestra que hasta los 110 000 €, no todas las concesiones presentan riesgos. Por último, para el importe total máximo de las concesiones

observadas (140 000 €), la probabilidad de sanciones se distribuye uniformemente entre 0,05 y 0,76. Esto es muy similar a lo que se observó en la distribución del número de concesiones: el número más alto se asocia a una distribución uniforme de los riesgos.

Figura 2.11. Distribución del importe total de las concesiones recibidas por el mismo tercero sobre la probabilidad de sanciones



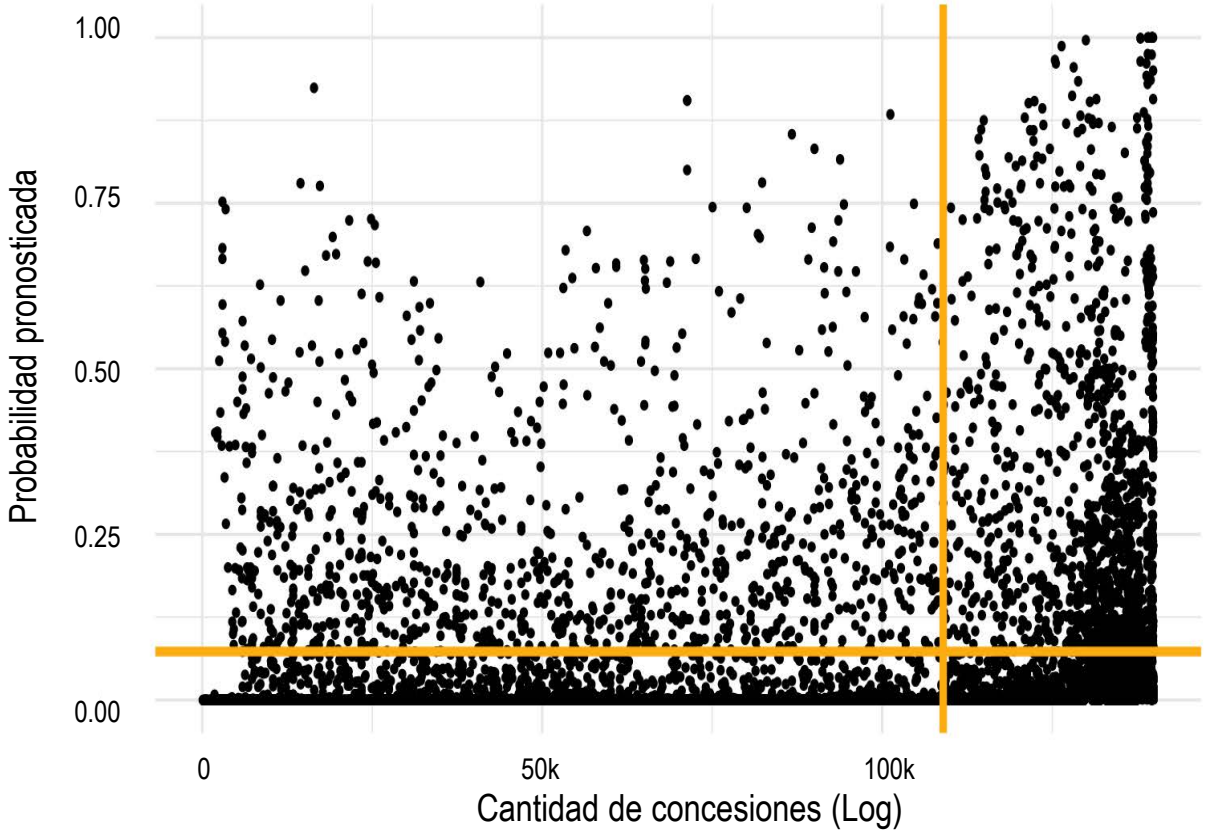
Fuente: Autor

Combinar las puntuaciones de riesgo pronosticadas con la información financiera

Los riesgos de fraude representan la variable clave de interés para IGAE y, por tanto, sirven como la principal variable dependiente para el modelo descrito hasta ahora. Sin embargo, solo representan una de las dimensiones claves según las cuales se pueden seleccionar los objetivos de la investigación. Una

segunda dimensión clave que podría tenerse en cuenta es el valor total de la subvención como una indicación del posible impacto económico del fraude para el gobierno español. La combinación de las puntuaciones de riesgo de fraude estimadas con el valor total de la concesión permite a quienes toman las decisiones y a los investigadores tener en cuenta, simultáneamente, la prevalencia de los riesgos y sus probables consecuencias económicas. (Fazekas, M., Ugale, G, & Zhao, A., 2019[7]). El enfoque más sencillo para observar estas 2 dimensiones simultáneamente es dibujar un diagrama de dispersión con estas 2 variables destacando también sus valores medios (Figura 2.12). El cuadrante superior derecho incluye aquellas concesiones que no solo tienen un alto riesgo, sino que también tienen valores altos de concesiones. Este es el grupo de mayor interés para las futuras investigaciones de la IGAE, ya que es más probable que incluyan subvenciones fraudulentas con importantes consecuencias económicas.

Figura 2.12. Distribución de concesiones según la puntuación de riesgo pronosticado y el valor total de la concesión



Fuente: Autor

Establecer un conjunto de datos preparado para detectar riesgos de fraude en el futuro

Para seguir mejorando el marco de evaluación de riesgos de fraude basado en datos de la IGAE, se pueden implantar una serie de reformas a corto y medio plazo que mejoren la calidad y el alcance de los datos subyacentes a los modelos de riesgo. El desarrollo de un modelo de riesgo por parte de la OCDE ya ha ayudado a la IGAE a avanzar en el tratamiento de algunos de estos problemas, y el conjunto de datos resultante del trabajo de la OCDE puede ser un punto de partida para la IGAE. No obstante, los

conjuntos de datos no son estáticos y pueden aparecer disponibles nuevas fuentes de datos. Por tanto, estos puntos son relevantes fuera del contexto de este proyecto.

Primero, los datos existentes pueden combinarse mejor y más rápido en un solo conjunto de datos preparado para modelar el riesgo de fraude. Actualmente, casi todos los conjuntos de datos abarcan distintas unidades de análisis como concesiones, convocatorias u organizaciones. Para que la IGAE los fusione, cada conjunto de datos debe alinearse al mismo nivel con ID únicas para evitar la multiplicación redundante de observaciones en el conjunto de datos fusionado. Durante el tratamiento de datos para este informe, estos se han transformado de formato largo a ancho cuando ha sido necesario. Sin embargo, este enfoque adolece de un gran inconveniente, que es un índice alto de omisión de ID sin múltiples observaciones por ID única. Para resolver este problema, se necesita la agregación, especialmente para las variables factoriales que no pueden calcularse como medias o medianas.

En segundo lugar, es fundamental reducir las tasas de omisión en todas las variables recopiladas por la IGAE. Como se trató en el Capítulo 1, definir estándares de calidad de datos y aplicarlos, en colaboración con los propietarios de datos, aseguraría que no haya variables con tasas de omisión elevadas, como 40 %-50 %. En tercer lugar, algunos conjuntos de datos (por ejemplo, sobre proyectos) consisten en una cantidad muy pequeña de observaciones, lo que impide su análisis junto con el conjunto de datos principal (es decir, cuando se combinan, dan como resultado un índice de omisiones alto). Del mismo modo, los datos sobre beneficiarios son muy limitados y necesitan mejoras adicionales.

Ampliar el uso de indicadores por parte de la IGAE a lo largo del ciclo de subvenciones.

Como se señaló, la lista final de indicadores incluye 29 variables. Estas variables son en su mayoría categóricas, aunque existen algunas variables numéricas, como importes y pagos. La mayoría de las variables analizadas son descriptivas debido a los datos disponibles. Hay datos limitados que pueden proporcionar información sobre los comportamientos de organizaciones y personas, como los conflictos de intereses entre los actores que reciben o se benefician de las subvenciones. Esta es una de las mayores lagunas en los datos actuales disponibles para la IGAE y es uno de los factores más restrictivos en su análisis de riesgo, independientemente de la metodología utilizada. La Tabla 2.5 muestra indicadores de comportamiento adicionales que podrían usarse para la evaluación de riesgos de fraude que abarcan el ciclo de la subvención y podrían ayudar a refinar el modelo de riesgo de la IGAE.

Tabla 2.5. Indicadores de comportamiento para evaluar los riesgos de fraude en cada fase del ciclo de subvenciones

Grupo de indicadores	Nombre del indicador	Definición del indicador
Fase de competición	falta de competición	Solo un solicitante para una convocatoria de subvenciones
Fase de selección	concentración de importes	Cantidad y valor excesivos de pagos a un solo proveedor
	Influencia política	Solo a los beneficiarios vinculados al gobierno se les conceden sus solicitudes
Fase de ejecución de la subvención	sobrefinanciación	Salario u otra compensación por servicios personales que exceden los importes aprobados por la agencia o son más altos que la compensación por otros servicios comparables que no están financiados por subvenciones.
	grandes pagos anticipados	Un beneficiario que extrae todos o la mayoría de los fondos de la subvención poco después de la concesión de la subvención puede ser característico de un riesgo más alto, a menos que el programa de subvenciones permita esta práctica.
	Modificación de plazos	Solicitud del contratista para modificar los plazos y las condiciones del contrato.
	Operación grande	Una sola operación representa más de la mitad de los costes totales del proyecto.
	Gastos atrasados	Gastos fuera del período permitido del proyecto
	Operaciones inusuales	Las operaciones cuestionables o inusuales inmediatamente anteriores al final

Grupo de indicadores	Nombre del indicador	Definición del indicador
		del período de concesión de una subvención pueden indicar que los defraudadores esperaron hasta el final del proyecto para retirar los fondos de la subvención para cubrir los costes no permitidos.
Organización destinataria	empresa nueva	Beneficiario final constituido inmediatamente antes de la solicitud de la subvención
	doble financiación	Prueba de que los beneficiarios están financiando proyectos de subvenciones con más de una subvención
	Viabilidad económica	Un destinatario que tiene una viabilidad financiera cuestionable, como un alto porcentaje de activos financiados con deuda o una liquidez insuficiente.
Contratistas y asesores	adquisición no competitiva	Destinatarios que gastan fondos en compras no aprobadas o subcontrataciones de adquisición de fuente única no aprobados o sin licitación
	subcontrataciones de asesores	El uso de asesores genéricos, no específicos o confusos
	documentación insuficiente	Justificación y documentación insuficientes para pagos realizados a contratistas/asesores, como horas trabajadas y actividades
Seguimiento y auditorías	consultas de auditoría	Varias consultas de las fuerzas del orden o las oficinas de auditoría que no pueden responderse
	no cooperación con auditores	El personal receptor que no coopera con las actividades de seguimiento o es agresivo con los auditores o gestores de subvenciones

Fuente: Autor

Existe una variedad de fuentes y ejemplos que pueden ayudar a la IGAE a mejorar sus indicadores de riesgo. En la Unión Europea, la Oficina Europea de Lucha contra el Fraude (OLAF) creó en 2011 un Compendio de casos anonimizados, que todavía tiene relevancia en la actualidad. El Compendio enumera los resultados de las investigaciones de la OLAF e incluye información sobre fraudes financieros. Se pueden identificar dos fases de alto riesgo de comportamiento fraudulento potencial: la fase de selección y la fase de ejecución. Durante la fase de selección, la OLAF fomentó que se inspeccionaran de cerca declaraciones justificativas y documentación oficial, así como a asegurarse de que el beneficiario final no se haya constituido o creado inmediatamente antes de la publicación de la subvención. Durante la fase de ejecución, la OLAF sugirió tener en cuenta las dificultades financieras del contratista, la presencia de una única operación grande que cubra casi la mitad de todos los costes del proyecto, así como el uso de la subvención para otros fines (European Anti-Fraud Office (OLAF), 2011[8]). El Compendio ilustra la realidad de que muchos fraudes son simplemente versiones recicladas de sistemas similares. De hecho, en su *32º Informe anual sobre la protección de los intereses financieros de la Unión Europea - Lucha contra el fraude -2020*, la Comisión Europea señaló que, entre las irregularidades fraudulentas relacionadas con la infraestructura sanitaria y la pandemia de COVID-19, los problemas más frecuentemente detectados se referían a la documentación de apoyo (European Commission, 2020[9]). El Recuadro 2.2 proporciona información adicional a partir de la experiencia del Comité de Fraude en Subvenciones del Grupo de Trabajo de Lucha contra el Fraude Financiero, que se creó para abordar el fraude a raíz de la crisis económica de 2008.

Recuadro 2.2. El Comité de Fraude en Subvenciones del Grupo de Trabajo de Lucha contra el Fraude Financiero de EE. UU.

En Estados Unidos, el Comité de Fraude en Subvenciones del Grupo de Trabajo de Lucha contra el Fraude Financiero identificó varias áreas clave para monitorizar e identificar actividades fraudulentas:

- la estructura de la organización destinataria y el programa de subvenciones
- solicitudes de pago o retiradas de fondos ordinarios
- seguimiento de informes y actividades

- actividades a nivel de operaciones
- contratos y asesores.

Entre la primera categoría, el Comité de Fraude en Subvenciones sugirió monitorizar el diseño del proyecto, así como la viabilidad económica del destinatario, el control interno, el personal de las organizaciones y los posibles conflictos de interés. En lo que respecta a las solicitudes de pago, se debe prestar atención al momento de la concesión de la subvención, así como a la documentación justificativa, al exceso de gastos y al redondeo de las cifras para la concesión de la subvención. Al realizar las actividades de seguimiento, la capacidad de respuesta y la cooperación del beneficiario son indicadores clave, así como la presencia de controles internos y el historial de auditorías de la empresa. Cuando se trata de actividades a nivel de operaciones, las operaciones excesivas, inusuales y no supervisadas podrían marcarse como riesgos potenciales y también como doble financiación (más de una subvención que cubra el mismo proyecto). Por último, en lo que respecta a los contratos y asesores, el Comité de Fraude en Subvenciones sugiere examinar las operaciones de partes relacionadas, el gasto en asesores no específicos y los beneficiarios de subvenciones con deficiencias en sus sistemas de adquisiciones. En caso de monitorización de datos, el Comité de Fraude de Subvenciones (2012) identifica los siguientes riesgos de fraude:

- número e importe excesivos de pagos a un solo proveedor
- pagos a proveedores no aprobados
- operaciones que eluden los procedimientos de revisión normales o que, de otro modo, no los controla otra persona
- compras que parecen irracionales, teniendo en cuenta la naturaleza del programa de subvenciones
- gastos fuera del período permitido del proyecto
- pagos y operaciones recurrentes
- pagos emitidos a varios proveedores a la misma dirección postal

Nota: En 2018, el Grupo de Trabajo sobre Integridad del Mercado y Fraude al Consumidor sustituyó al Grupo de Trabajo de Ejecución de Fraude Financiero.
Fuente: (Financial Fraud Enforcement Task Force, 2012[10])

Invertir en la mejora continua del modelo de riesgo

Dado que la validez de este modelo basado en datos depende de las sanciones impuestas, si las sanciones no cumplieran con los esquemas de fraude relevantes o resultaran en una muestra sesgada de investigaciones, cualquier modelo de evaluación de riesgos también estaría sesgado. Por tanto, obtener una muestra verdaderamente aleatoria de investigaciones y sanciones es de vital importancia. Con este fin, la IGAE puede seleccionar un porcentaje de los casos investigados cada año utilizando sus técnicas tradicionales de muestreo o un método de selección basado en datos como el presentado anteriormente. El resto de los casos investigados puede elegirse mediante una selección aleatoria completa. Este enfoque lograría un equilibrio entre optimizar la utilidad de los recursos de investigación a través de una mejor focalización, mientras que también invertiría en futuras mejoras del modelo de evaluación de riesgos, al proporcionar una muestra de entrenamiento mejor. También le daría a la IGAE una mejor idea del desempeño del modelo. Como este esfuerzo presente ha sido una prueba conceptual, se pueden valorar pasos técnicos adicionales:

- Mejorar la calidad de las variables en el todo el conjunto de datos para poder incluir más indicadores en el modelo, y por tanto mejorar su calidad.

- Se debe tener en cuenta la naturaleza desequilibrada de las clases en la variable dependiente (positiva/negativa): utilizar técnicas de insaculación PU para evitar imprecisiones en el modelado.
- Repetir el ejercicio de modelización con regularidad a medida que se disponga de nuevos datos, incluidas las sanciones y las concesiones, para mantener actualizada la evaluación de riesgos.
- Los modelos analíticos no dibujan una imagen completa y pueden tener sesgos a medida que aprenden de las acciones de control pasadas (ver Capítulo 1). La IGAE podría complementar los modelos con métodos cualitativos y el criterio experto. Esto permite a los especialistas en fraude de la IGAE contribuir con su conocimiento especializado sobre sistemas de fraude, los últimos eventos y el contexto más amplio.

Si bien los modelos pueden ser vulnerables a los sesgos en sí mismos, también pueden ayudar a controlarlos. Concretamente, la selección de muestras basada en datos, incluidas las que utilizan el aprendizaje automático, no solo ayudaría a la IGAE a optimizar la eficacia de los recursos de control, sino que también ayudaría a corregir algunos sesgos en el conjunto de datos de entrenamiento. Por ejemplo, si se conocen los tipos de fraude que no están cubiertos por las investigaciones, sus características se pueden introducir manualmente en la base de datos para proporcionar información suficiente para que el algoritmo aprenda. Además, si la selección de investigaciones en el conjunto de sanciones enfatiza ciertas variables, como el importe de la subvención, submuestrear las subvenciones grandes y las subvenciones pequeñas puede contrarrestar la selección sesgada de los casos investigados.

Tener en cuenta los análisis de redes y utilizar un conjunto más amplio de metodologías

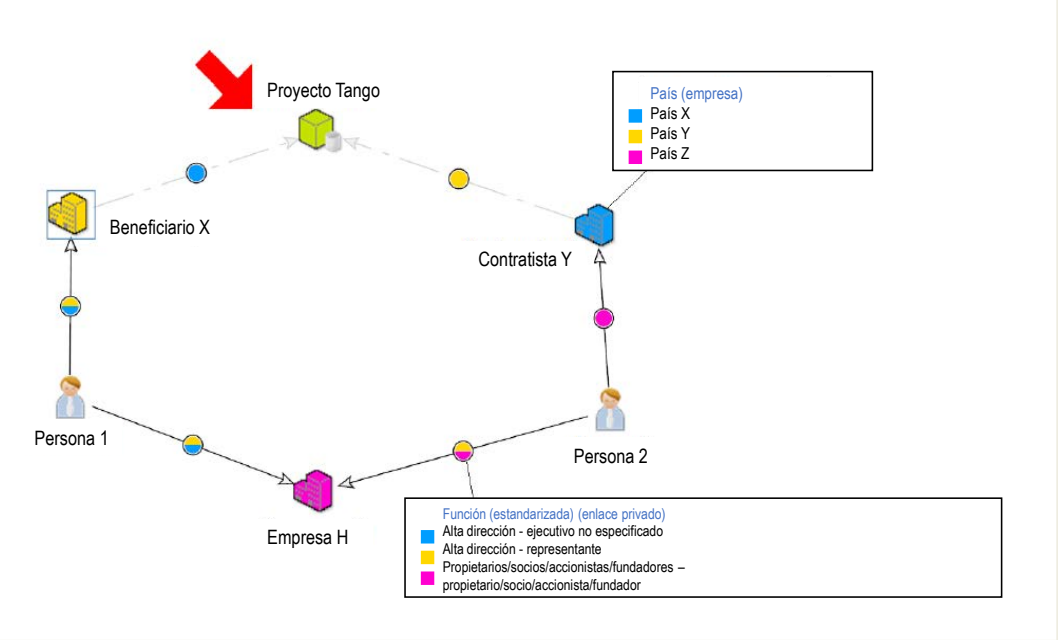
Las técnicas de ciencia de datos y redes se han utilizado cada vez más para estudiar los delitos económicos como la corrupción, el fraude, la colusión, la delincuencia organizada o la evasión fiscal, por mencionar algunas áreas importantes (Wachs, Fazekas and Kertész, 2020[11]). Explorar redes sin análisis avanzado ya promete grandes ventajas para la detección de fraudes, como el rastreo de posibles conflictos de intereses (ver el Recuadro 2.3).

Recuadro 2.3. Usar datos para investigar conflictos de interés

Cuando se conocen las personas que están detrás de las organizaciones públicas y privadas que participan en el proceso de concesión de la subvención, se puede descubrir una serie de posibles de relaciones con conflicto de interés. Si bien las investigaciones en profundidad pueden revelar dichas relaciones, la detección de riesgos se facilita en gran medida al hacer coincidir conjuntos de datos a gran escala que contienen: 1) todos los cargos públicos que desempeñan un papel importante en la preparación, evaluación, concesión y seguimiento de subvenciones; y 2) todas las personas físicas que tienen un papel importante en las empresas que presentan solicitudes, reciben y ejecutan subvenciones.

La recopilación, limpieza y vinculación de dichos conjuntos de datos y el mantenimiento de las conexiones subyacentes pueden generar costes importantes. Sin embargo, una vez que se dispone de un conjunto de datos de este tipo y una interfaz gráfica simple, como es el caso de la herramienta ARACHNE de la UE, se puede acelerar enormemente la selección e investigación de las relaciones de riesgo entre los concedentes y los beneficiarios de las mismas. Por ejemplo, es posible examinar de forma rápida y eficiente los proyectos, los beneficiarios y las personas que participan en la preparación de la convocatoria y la evaluación de las solicitudes.

Figura 2.13. Visualización de conflictos de intereses



Fuente: (Unión Europea, 2016[12])

El análisis a gran escala de redes de relaciones contractuales o personales puede revelar patrones ocultos que sirven como indicadores de riesgo por sí mismos o complementan otros indicadores de riesgo (Fazekas and Tóth, 2016[13]); (Fazekas and Wachs, 2020[14]). Por ejemplo, los indicadores de riesgo de licitación, como la ocurrencia de licitación única en la contratación pública, pueden superponerse a grupos de compradores y proveedores vinculados en la contratación pública para identificar grupos organizados de alto riesgo. La Figura 2.14 a continuación muestra una visualización de relaciones de compradores y proveedores en contratación pública en Hungría. Dichos diagramas proporcionan una instantánea visual de los datos que señalan las posibles relaciones de alto riesgo para una investigación ulterior. Por ejemplo, las líneas rojas resaltan un indicador de oferta única más alto que la media de ofertas únicas en esa relación. Además, hay un grupo de actores de alto riesgo de corrupción en la parte superior (es decir, relaciones densas de contratación que coinciden con índices altos de licitación única en esas relaciones).

Figura 2.14. Relaciones de compradores y proveedores en contratación pública, Hungría 2014

Fuente: (Wachs, Fazekas and Kertész, 2020[11])

La IGAE puede recopilar conjuntos de datos relevantes, como datos de propiedad de las empresas, y vincularlos a sus datos básicos de subvenciones, para hacer uso de dichas técnicas de análisis de redes. A medida que las personas se mueven entre los sectores público y privado y hay otras formas en que los beneficiarios pueden establecer conexiones con los organismos que conceden subvenciones, el rastreo de redes abiertas u ocultas ofrece una herramienta clave para mejorar la evaluación de riesgos de fraude en España. Como se ha tratado, esta es un área en la que la IGAE actualmente tiene lagunas en sus datos, por lo que el uso de análisis de redes también dependerá de la capacidad de la IGAE para abordar estas lagunas. Los acontecimientos recientes en España sugieren que ya se están realizando mejoras. Por ejemplo, en mayo de 2020, la IGAE y la Tesorería General de la Seguridad Social (TGSS) firmaron un convenio sobre transferencia de información, estableciendo condiciones más colaborativas para el control financiero de las subvenciones y ayudas públicas. El acuerdo estipula el acceso directo a las bases de datos de la TGSS para facilitar el trabajo de la IGAE en la detección de fraude e irregularidades (Ministry of the Presidency of Spain, 2021[15]). Avanzar en acuerdos similares con otras entidades públicas y privadas, concretamente para obtener datos empresariales y datos que reflejen indicadores de comportamiento, como se mencionó anteriormente, constituyen aportaciones fundamentales para fortalecer futuros modelos de riesgo.

Conclusión

Este capítulo ha presentado una prueba de concepto para que la IGAE mejore su enfoque de evaluación de los riesgos de fraude en las subvenciones públicas, basándose en las principales prácticas de análisis. El proceso de desarrollo del modelo de riesgo ha llevado a una serie de descubrimientos sobre la capacidad actual de análisis de la IGAE, así como de la gestión de datos y la garantía de calidad de los datos, con el fin de evaluar los riesgos de fraude, como se indica en el Capítulo 1. En el desarrollo del modelo de riesgo también se han encontrado lagunas en los indicadores y bases de datos de riesgo de fraude, que si se abordan pueden ayudar a la IGAE a mejorar sus evaluaciones de riesgo de fraude, independientemente de la metodología específica que elija. En particular, la IGAE podría incorporar indicadores de comportamiento adicionales para evaluar riesgos de fraude en cada fase del ciclo de subvenciones, basándose en experiencias internacionales y literatura académica. Además, la IGAE podría recopilar datos empresariales y adoptar las metodologías descritas para realizar análisis de redes como un medio para identificar conflictos de interés, basándose en los ejemplos de compras públicas de este capítulo.

El capítulo también recoge una serie de consideraciones técnicas para la IGAE si decide adoptar un modelo de aprendizaje automático. Gran parte del trabajo pesado se ha realizado como parte de este proyecto piloto en términos de procesamiento y limpieza de datos. La IGAE tiene ahora un conjunto de datos de trabajo para usar en el análisis de riesgo de fraude, que ya es una mejora de lo que tenía disponible antes de este proyecto. Las características y limitaciones de los datos de la IGAE motivaron en gran parte la justificación para seleccionar el enfoque descrito. Si bien tiene limitaciones debido a la calidad de los conjuntos de datos de entrenamiento y las características de los datos de las subvenciones, la metodología se diseñó para minimizar los falsos positivos y los falsos negativos y, en general tiene un alto poder predictivo para identificar posibles fraudes en las subvenciones públicas de España. Si bien la implantación de este enfoque requiere capacidades adicionales, que se detallan en el Capítulo 1, la prueba de concepto muestra con éxito lo que es posible hacer con una inversión modesta, y proporciona una base para que la IGAE adopte una evaluación de riesgo de fraude verdaderamente basada en datos. El Capítulo 3 indaga más en cómo la IGAE puede mejorar la precisión del modelo interpretando datos adicionales que pueden usarse para detectar posibles fraudes.

Referencias

Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45/1, pp. 5-32, https://link.springer.com/article/10.1023/a:1010933404324. [1]

Elkan, C. and K. Noto (2008), "Learning classifiers from only positive and unlabeled data", *KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213-220, https://dl.acm.org/doi/10.1145/1401890.1401920. [2]

European Anti-Fraud Office (OLAF) (2011), *Compendium of Anonymised Cases*, https://ec.europa.eu/sfc/sites/default/files/sfc-files/OLAF-Intern-2011.pdf (accessed on 13 August 2021). [8]

European Commission (2020), *Report from the Commission to the European Parliament and the Council: 31st Annual Report on the protection of the European Union's financial interests — Fight against fraud - 2019*, https://ec.europa.eu/anti-fraud/sites/default/files/pif_report_2019_en.pdf (accessed on 13 August 2021). [9]

Fazekas, M., Ugale, G, & Zhao, A. (2019), *Analytics or Integrity: Data-Driven Decisions for Enhancing Corruption and Fraud Risk Assessments*, OECD Publishing, Paris, https://www.oecd.org/gov/ethics/analytics-for-integrity.pdf. [7]

Fazekas, M. and I. Tóth (2016), "From Corruption to State Capture: A New Analytical Framework with Empirical Applications from Hungary", *Political Research Quarterly*, Vol. 69/2, pp. 320-334, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=H1FpS2AAAAAJ&citation_for_view=H1FpS2AAAAAJ:SP6oXDckpogC. [13]

Fazekas, M. and J. Wachs (2020), "Corruption and the network structure of public contracting markets across government change", *Politics and Governance*, Vol. 8/2, pp. 153-166, https://scholar.google.fr/citations?view_op=view_citation&hl=fr&user=PY3YH2kAAAAJ&citation_for_view=PY3YH2kAAAAJ:ZeXyd9-uunAC. [14]

Financial Fraud Enforcement Task Force (2012), *Reducing Grant Fraud Risk: A Framework For Grant Training*, https://www.oversight.gov/sites/default/files/oig-reports/Grant-Fraud-Training-Framework.pdf. [10]

James, G. et al. (2015), *Chapter 8*, https://link.springer.com/book/10.1007/978-1-4614-7138-7. [5]

Li, C. and X. Hua (2014), "Towards positive unlabeled learning", *International Conference on Advanced Data Mining and*, pp. 573–587. [3]

Lundberg, S. and S. Lee (2017), *A unified approach to interpreting model predictions*, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=ESRugcEAAAAJ&citation_for_view=ESRugcEAAAAJ:dfsIfKJdRG4C. [6]

Ministry of the Presidency of Spain (2021), *Resolution of May 26, 2020, of the Undersecretariat, which publishes the Agreement between the General Treasury of Social Security and the General Intervention of the State Administration, on the transfer of information*, https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-5748 (accessed on 4 July 2021). [15]

Mordelet, F. and J. Vert (2014), “A bagging SVM to learn from positive and unlabeled examples.”, *Pattern Recognition Letters*, Vol. 37, pp. 201-209, https://www.sciencedirect.com/science/article/pii/S0167865513002432. [4]

Unión Europea (2016), *Arachne, Be Distinctive*, http://www.ec.europa.eu/social/BlobServlet?docId=15317&langId=en (accessed on 13 August 2021). [12]

Wachs, J., M. Fazekas and J. Kertész (2020), “Corruption risk in contracting markets: a network science perspective”, *International Journal of Data Science and Analytics*, pp. 1-16, https://scholar.google.fr/citations?view_op=view_citation&hl=fr&user=PY3YH2kAAAAJ&citation_for_view=PY3YH2kAAAAJ:QIV2ME_5wuYC. [11]

Notas

[1] Para limpiar y fusionar los datos, se utilizó R 3.6.3 con el siguiente paquete: readxl, tidyverse (dplyr), flipTime, tibble, data.table. Para el modelado, se utilizaron R 3.6.3 y Python3 para las distintas fases de análisis. Para construir random forests en R, se utilizaron los paquetes randomForest y xgboost. Para el aprendizaje positivo sin etiquetar en las bibliotecas de Python3, se utilizaron pandas, numpy, baggingPU (módulo BaggingClassifierPU), sklearn.tree (módulos DecisionTreeClassifier, DecisionTreeRegressor, precision_score, recall_score, precision_score, train_test_split, RandomForestClassifier).

[2] Los nombres de estos conjuntos de datos incluyen BDNS_CONV_ACTIVIDADES, BDNS_CONV_ANUNCIOS, DNS_CONV_FONDOS_CON690, BDNS_CONV_OBJETIVOS_CON503, BDNS_CONV_TIPOBEN_CON590, BDNS_CONV_REGIONES_CON570, BDNS_CONV_INTRUMENTOS_CON560.

[3] Los nombres de estos conjuntos de datos incluyen BDNS_PROYECTOS, BDNS_PAGOS, BDNS_REINTEGRO, BDNS_DEVOLUCIONES.

[4] Los nombres de estos conjuntos de datos incluyen BDNS_INHABILITACIONES, BDNS_SANCIONES y BDNS_TERCERO_ACTIVIDADES_TER320

[5] El índice de precisión tan alto (95 %) se debe en gran parte al hecho de que la muestra está desequilibrada; es decir, la mayoría de los casos son negativos (no sancionados) y, por tanto, el modelo puede clasificar con relativa facilidad la mayor parte de la muestra como no sancionada. Sin embargo, es más difícil para el modelo predecir correctamente los casos sancionados, dado que son mucho menos frecuentes. Para este caso, la calificación de repetición es más útil para evaluar el rendimiento del modelo, ya que calcula el número de miembros de una clase que el clasificador identificó correctamente dividido por el número total de miembros de esa clase.

3 Mirando hacia el futuro: Un mapa de ruta de conjuntos de datos para mejorar el modelo de riesgo de fraude de la Intervención General de España

Este capítulo explora conjuntos de datos adicionales que la Intervención General de la Administración del Estado (IGAE) de España puede utilizar para mejorar el modelo de riesgo descrito en el Capítulo 2. El capítulo proporciona una hoja de ruta e indica qué bases de datos son más prometedoras para mejorar la evaluación de riesgo de fraude de subvenciones utilizando el modelo, en función de la accesibilidad, relevancia y calidad de los conjuntos de datos. Los conjuntos de datos se agrupan en tres categorías: 1) datos organizativos de las partes del proceso de concesión; 2) datos sobre conexiones personales y conflictos de intereses; y 3) datos sobre fiabilidad organizativa e infracción de normas.

Introducción

Este capítulo presenta un mapa de ruta para complementar los datos de subvenciones existentes de la Intervención General de la Administración del Estado (IGAE), con el fin de mejorar los modelos de evaluación de riesgos. Implícitamente, describe conjuntos de datos que pueden enlazar con datos ya existentes de subvenciones de la IGAE, mejorando así la sofisticación analítica y la precisión de la evaluación de riesgos de la IGAE. Como se trató en el Capítulo 2, los modelos de aprendizaje automático están limitados por el alcance y el tipo de datos incluidos en la muestra de entrenamiento. El modelo no puede estimar con precisión las probabilidades de riesgo basándose en información incompleta, porque hay factores claves y mecanismos que determinan los riesgos siguen sin ser considerados. Por tanto, cuanto más completo es el conjunto de datos inicial, más precisos y exactos se vuelven los cálculos de riesgos.

Dado que el universo de conjuntos de datos potencialmente relevantes es enorme, es imperativo reducir la lista de a los más relevantes antes de invertir recursos importantes en el mapeo, tratamiento, vinculación y posterior incorporación de datos a los modelos predictivos. Se deben tener en cuenta tres factores al seleccionar conjuntos de datos adecuados: *Accesibilidad, relevancia y calidad.* La *accesibilidad* en este contexto es la facilidad con la que el conjunto de datos se puede obtener de su fuente original, lo que puede incluir cuestiones como si el conjunto de datos se puede descargar públicamente o debe solicitarse. El formato en el que están disponibles los datos también es fundamental, si existe como conjunto de datos descargable, o está constituido por una serie de páginas HTML. La *relevancia* se refiere al potencial de los elementos de datos para mejorar la sofisticación y precisión analítica. Esto debe evaluarse antes de recopilar los datos. La prueba definitiva de esta evaluación inicial es si los datos mejorarían la precisión predictiva del modelo. Cuando se incluyen demasiadas variables redundantes, el modelo final puede sufrir un sobreajuste. La *calidad* de los datos en este contexto captura la tasa de valores no omitidos y la fiabilidad de la información. Es probable que los datos de poca calidad con muchos valores omitidos o datos recopilados de manera inexacta sesguen los resultados. Este capítulo solo abarca los conjuntos de datos que se consideren fácilmente accesibles para la IGAE, relevantes para dicho modelo de riesgo y de suficiente calidad.

Mapa de ruta para complementar los datos de subvenciones de la IGAE

Los dos capítulos anteriores han descrito el proceso mediante el cual se puede implantar el aprendizaje automático para mejorar el enfoque de la IGAE para identificar riesgos en la concesión de subvenciones y ayudas. El proceso de utilizar conjuntos de datos externos, además de los datos internos existentes, sigue la misma lógica. Primero, se deben definir indicadores de antecedentes y de riesgo para cada conjunto de datos, para identificar los factores que potencialmente influyen en los riesgos de fraude. El siguiente paso es vincular los conjuntos de datos al conjunto de datos interno ya existente. Para hacerlo, se deben tener en cuenta algunas cosas: La unidad de análisis en cada conjunto de datos, la relevancia de la variable, el índice de omisión y la varianza. Como se trataba en el Capítulo 2, el índice de omisión debe ser inferior al 50 %, con una varianza de al menos el 35 %. Además, para fusionar nuevos datos, deben alinearse con la misma unidad de análisis con ID únicas, para evitar filas duplicadas después de la combinación. Hay que descartar las variables que no contienen información útil (es decir que no pueden utilizarse como indicadores).

Por ejemplo, para agregar conjuntos de datos externos a la Base de Datos Nacional de Subvenciones (BDNS) existente, deben tener identificaciones que coincidan con los utilizados en BDNS. Dichas identificaciones incluyen identificaciones de subvenciones, Número de Identificación Fiscal (NIF) español de los beneficiarios y nombres de los órganos concedentes, así como los nombres de los municipios. Esto implica algunas limitaciones. Por ejemplo, actualmente es imposible hacer combinar beneficiarios por sus nombres; solo pueden combinarse por NIF. Además, combinar por municipio acarreará una pérdida

significativa de datos, porque alinear los datos con la misma unidad de análisis con identificaciones únicas significa que las puntuaciones de riesgo deben agregarse por municipio. Se aplica una lógica similar a la combinación por nombres de los concedentes y el NIF de los beneficiarios, ya que hay muchos valores idénticos en todos los datos de la BDNS (es decir, el mismo beneficiario puede recibir varias subvenciones o ayudas).

Hay algunas fuentes, unas más fiables que otras, que pueden usarse potencialmente para agregar datos al conjunto de datos existente de la BDNS. En primer lugar, están las fuentes oficiales, como el Registro Nacional de Asociaciones del Ministerio del Interior, que registra las organizaciones no gubernamentales (ONG) acreditadas, la base de datos tributaria de la Agencia Estatal de Administración Tributaria (AEAT) y la Asociación Española de Fundaciones (listas fundaciones acreditadas). Algunos de los datos son de acceso público, mientras que otros están restringidos solo a agencias autorizadas.

Los registros de propiedad efectiva (BO) y los datos de contratación pública también pueden considerarse fuentes oficiales fiables. La ventaja de trabajar con datos oficiales obtenidos directamente de los titulares de datos es que no es necesario verificar la información facilitada, más allá de las verificaciones de calidad de datos estándar utilizadas como parte de las canalizaciones de datos descritas. Los datos oficiales sobre ayudas de la Unión Europea son otro ejemplo de datos fiables.

El siguiente grupo de fuentes son las ONG y asociaciones independientes. Esta información es menos fiable, ya que el proceso de recopilación y verificación de datos no es claro. Si bien las fuentes oficiales probablemente incluyen datos e información primarios, las fuentes secundarias se adquieren de diferentes fuentes o se recopilan manualmente, a menudo sin transparencia sobre cómo se forma el conjunto de datos. Por tanto, estos conjuntos de datos deben usarse con más cuidado y su validez debe verificarse más a fondo. En España, entre dichas fuentes, se encuentran evaluadores independientes de ONG, así como FICESA, una base de datos de altos cargos y directivos.

Descripción general de los grupos de conjuntos de datos más relevantes

Hay cuatro grupos principales de datos que son relevantes para vincular con la base de datos BDNS, con el fin de mejorar las evaluaciones de riesgo de fraude de la IGAE. Cada grupo puede facilitar información sobre dimensiones y factores claramente distintos de riesgos de fraude. Algunos datos crean oportunidades para métodos alternativos de análisis, como el análisis de redes, que revelan conexiones entre empresas privadas y personas políticamente expuestas, así como beneficiarios finales y empresas asociadas. Reunir todos estos conjuntos de datos ofrece la oportunidad de realizar la evaluación de riesgos más completa; sin embargo, hacer coincidir solo algunos, o incluso solo un conjunto de datos adicional, puede ser muy útil para mejorar el modelo de riesgo de la IGAE, incluidos los siguientes grupos de datos:

i. **Datos organizativos de las partes del proceso de concesión.** Este grupo abarca datos sobre concedentes y beneficiarios, así como sobre terceros (como implantadores de proyectos). Las posibles fuentes de información para este grupo son:

- Registro de empresas e información financiera: proporciona información sobre la estructura organizativa y la historia de la empresa (por ejemplo, cuándo se fundó) y también descubre la situación financiera, como la rentabilidad de la organización.
- Datos organizativos sobre ONG acreditadas, fundaciones, asociaciones: proporciona información sobre las características del registro, la fiabilidad de la organización y los registros financieros.

ii. **Datos sobre conexiones personales y conflictos de interés.** Este grupo puede ser útil para identificar conexiones entre cargos de organizaciones privadas que solicitan subvenciones y los responsables políticos que supervisan las subvenciones. Conectar cargos públicos y privados puede resultar útil para

seguir investigando posibles conflictos de intereses. Las posibles fuentes de información para este grupo son:

- El registro de propiedad efectiva (BO): puede ayudar a identificar a los beneficiarios finales, las empresas asociadas y sus registros.
- Personas políticamente expuestas: ayuda a revelar a las personas a las que se les ha confiado el poder y que son más susceptibles de verse envueltas en sobornos u otras prácticas corruptas.
- Datos sobre altos cargos y directivos: proporciona nombres de personas potencialmente vinculadas a empresas privadas a través de propiedad legal o propiedad efectiva

iii. **Datos sobre fiabilidad de las organizaciones e infracción de normas.** Este grupo puede ayudar a predecir riesgos de fraude, al ofrecer información sobre infracciones relevantes, pero solo indirectamente relacionadas, como irregularidades en el pago de impuestos. Este grupo también puede proporcionar información sobre medidas de fiabilidad más suaves, como la acreditación ante la sociedad civil. Las posibles fuentes de información son:

- Datos sobre quiebras o pagos de impuestos: muestra la fiabilidad de una organización basada en registros financieros pasados:
- Acreditaciones de ONG: identifica a las ONG acreditadas u otras asociaciones como más fiables.

iv. **Datos sobre otros fondos y contratos.** La información sobre otras fuentes de financiación y contratos públicos puede revelar factores adicionales que influyen en la probabilidad de fraude, como la doble financiación para la misma actividad. Además, los riesgos de corrupción en contratación pública u otros procesos de financiación pueden apuntar debilidades sistemáticas a escala organizativa y la propensión a cometer fraude. Los conjuntos de datos relevantes en este grupo incluyen:

- Fondos de la UE: la lista de beneficiarios de ayudas de la UE puede mostrar si la organización recibió financiación doble de diferentes fuentes para el mismo proyecto.
- Contratación pública: los riesgos de corrupción en los contratos públicos obtenidos de organizaciones o proporcionados por el mismo concedente pueden influir en la posibilidad de irregularidades en subvenciones y ayudas.

La Tabla 3.1 presenta los conjuntos de datos más prometedores en España, que bien son de acceso público o que su contenido y características son de dominio público. Para cada conjunto de datos que pertenece a uno de los 4 grupos de conjuntos de datos, la tabla contiene información sobre la unidad de medida (que se refiere una sola fila), el número de observaciones cuando estén disponibles, la identificación clave para enlazar con la BDNS[1] y la prioridad para el futuro trabajo de la IGAE. La tabla destaca los conjuntos de datos de máxima prioridad en la parte superior, teniendo en cuenta las tres dimensiones principales de la evaluación de datos tratadas anteriormente: Accesibilidad, relevancia y calidad. Solo los conjuntos de datos que han obtenido puntuaciones altas en las 3 dimensiones (descarga masiva de datos de fácil *acceso*, alcance y contenido de datos muy *relevantes* y *calidad* adecuada) se han considerado de prioridad alta para la IGAE.

Por el contrario, algunos conjuntos de datos que han obtenido puntuaciones altas en solo una o dos dimensiones se han considerado de prioridad media o baja. Por ejemplo, cuando la accesibilidad a los datos es limitada, la prioridad se consideraba media incluso para los datos que de otro modo se hubieran considerado muy relevantes o de calidad adecuada. La clasificación de los conjuntos de datos en términos de prioridad general establece la ruta detallada para ampliar y enriquecer el conjunto de datos actual de la IGAE y el modelo de riesgo descrito en el Capítulo 2. Las siguientes secciones analizan cada uno de estos conjuntos de datos en detalle, junto con algunos indicadores de riesgo de fraude, que se pueden calcular cuando se conectan datos.

Tabla 3.1. Breve descripción de conjuntos de datos adicionales

Nombre del conjunto de datos	Grupo de conjuntos de datos	Unidad de medida	Número de observaciones	ID para combinar con el conjunto de datos principal de la IGAE	Prioridad para el trabajo de seguimiento de la IGAE
Registradores De España	i, ii	Organización	>5 000 000	NIF de beneficiarios, nombres de organizaciones	alta
Registro de propiedad efectiva (LibreBOR)	i, ii	Organización	>5 000 000	NIF de beneficiarios	alta
Base de datos de altos cargos y directivos españoles (FICESA)	ii	Instituciones y organismos estatales	~100 000	Nombre de organizaciones	alta
CINCOnet	iii	Organizaciones	debe acceder un organismo oficial	NIF de organizaciones	alta
Plataforma de contratación pública	iv	Licitación	1 391 558	NIF de organizaciones	alta
El Registro Público Concursal	iii	Organizaciones	el sitio web no permite buscar	NIF de organizaciones	media
La Asociación Española de Fundaciones (AEF)	iv	Base	15 840	Ubicación y tipo de beneficiario	media
Agencia Estatal de Administración Tributaria, AEAT	iii	Organizaciones	no en acceso público	NIF de organizaciones	media
Ayudas de la Unión Europea	iv	Subvención o contrato	40 567	Nombre del beneficiario, NIF	media
El Registro Nacional de Asociaciones	i, iii	ONG acreditada	44	NIF de la organización	baja
Fundación Lealtad	i, ii, iii	ONG acreditada	191	Nombre de la organización	baja

Fuente: Autor

Combinación de datos organizativos: perfiles organizativos más precisos y detección de anomalías

Los datos organizativos de las partes involucradas en la concesión de subvenciones incluyen los concedentes, los beneficiarios y los terceros (es decir, los ejecutores del proyecto). La combinación de datos sobre organizaciones permite obtener una visión más completa y detallada de los controles organizativos ante los riesgos de fraude. Ayuda a identificar características organizativas adicionales que podrían influir en la probabilidad de sanciones. Por ejemplo, la información contable, el tamaño de la empresa y las empresas asociadas pueden ser características útiles para identificar los riesgos de fraude y mejorar el modelo de riesgo de la IGAE en el futuro. Este grupo incluye las siguientes bases de datos: Registradores de España, datos de la Asociación Española de Fundaciones (AEF) y el Registro Nacional de Asociaciones del Ministerio del Interior.

Registro mercantil y datos financieros

Uno de los conjuntos de datos más relevantes para el propósito de la IGAE y para mejorar el modelo de riesgo es el Registro Mercantil nacional. Contiene datos sobre las empresas, el capital, los representantes

(por ejemplo, consejeros y abogados), los actos registrados y la presentación de cuentas anuales (es decir, el desempeño financiero). La lista de variables se presenta en la Tabla 3.2.[2]

Tabla 3.2. Lista de variables (Registro Mercantil Nacional)

Variables	Descripción	Tipo de variable
Nombre	El nombre de la empresa	Texto
NIF	El NIF de la empresa	Texto
Fecha de creación	La fecha de constitución de la empresa	Fecha
Domicilio social	La dirección en la que está registrada la empresa	Texto
Sector de actividad económica	En qué sector económico opera la empresa (NACE)	Categórico
Forma jurídica	Forma jurídica oficial de la empresa (formas nacionales)	Categórico
Estado de la empresa	Si la empresa está activa y operativa	Categórico
Activos de la empresa	Valor total de los artículos que benefician económicamente a la empresa	Numérico
Pasivos de la empresa	Valor total de las obligaciones de la empresa	Numérico
Ingresos de la empresa	Cantidad total de ingresos generados anualmente	Numérico
Gastos de la empresa	Importe total de gastos al año	Numérico
Cambios en el patrimonio	Si hubo cambios en el patrimonio neto durante el año anterior	Binario + texto
Liquidez	Incremento o disminución de la cantidad de dinero	Lista
Miembros	Incluye el nombre de todos los miembros de la representación empresarial actual	Texto
Propietarios efectivos	Lista de nombres de los propietarios finales de la empresa.	Texto

Fuente: https://sede.registradores.org/site/home

El Registro Mercantil puede cruzarse con el conjunto de datos principal de la BDNS por el NIF de la empresa, o en caso de error, por el nombre de la organización. Casi todos los elementos de datos que contiene el conjunto de datos de empresas son relevantes para la IGAE, en lo referente a mejorar su modelo de riesgo. Estos campos van desde la información básica de registro, como la fecha de creación o el domicilio social, hasta los balances y estados de resultados. Del mismo modo, las variaciones recientes en el patrimonio y la lista completa de accionistas de la empresa pueden proporcionar información adicional sobre posibles conflictos de interés cuando se cruzan con otros conjuntos de datos.

Con respecto a la información básica del registro, existen señales de alerta que han demostrado ser útiles para predecir los riesgos de corrupción y fraude. Por ejemplo, las empresas que se han constituido, o cuyos datos de registro se han modificado poco antes de solicitar una subvención, tienen un riesgo mayor. Del mismo modo, las empresas registradas en las llamadas direcciones de «cementerio de empresas» pueden ser de alto riesgo, donde un gran número de empresas están registradas con altos grados de fluctuación (por ejemplo, miles de empresas creadas y cerradas con la misma sede social cada mes). De manera similar, como se trata en el Capítulo 2, el tipo de organización (es decir, la naturaleza jurídica de la empresa), así como sus ingresos y tamaño generales, pueden influir en el nivel de riesgo de fraude. Por ejemplo, debido a la legislación, ciertos tipos de organizaciones pueden ser menos transparentes o estar menos reguladas (por ejemplo, fideicomisos o propiedad empresarial presentada por acciones al portador).

En cuanto a los datos financieros de la empresa, la IGAE puede tener en cuenta una serie de indicadores relevantes para la predicción de riesgos. Primero, la relación entre gastos e ingresos de una empresa puede proporcionar información sobre si la empresa es rentable. Las empresas que no son rentables son de mayor riesgo en subvenciones y ayudas, ya que pueden utilizar los fondos para pagar sus deudas en lugar de financiar sus proyectos. Del mismo modo, una relación negativa entre los pasivos y los activos de una empresa sugiere un mayor riesgo en términos del uso adecuado de las subvenciones. Los cambios frecuentes en el capital social pueden ser una señal de conflictos internos e inestabilidad dentro de la empresa, lo que aumenta el nivel de riesgo asociado a subvenciones y ayudas para dichas

organizaciones. La disminución sistemática de liquidez refleja el estancamiento o la reducción de la actividad de la empresa, lo que también pone en tela de juicio su viabilidad. Combinar los datos de las subvenciones con los datos financieros de la empresa también puede revelar el tamaño relativo de la subvención en comparación con la empresa, ya que las pequeñas empresas que reciben subvenciones importantes pueden ser de riesgo.

Registro de Asociaciones

Otro conjunto de datos organizativos que la IGAE podría tener en cuenta para su modelo de riesgo, aunque de baja prioridad, es el **Registro Nacional de Asociaciones**, del Ministerio del Interior. Se trata de un listado de organizaciones que han pasado una revisión realizada por la Agencia Española de Cooperación Internacional para el Desarrollo (AECID), en la que se utilizaron más de 70 criterios cualitativos y cuantitativos, en su mayoría relacionados con la experiencia, solvencia económica, transparencia y recursos humanos. La principal limitación de este conjunto de datos es el pequeño número de ONG acreditadas que proporciona, ya que solo tiene 44 entidades. Se almacenan en formato HTML y se pueden exportar fácilmente a Excel o cualquier otro formato de datos. El directorio de variables se detalla en la Tabla 3.3.

Tabla 3.3. Directorio de variables (Registro Nacional de Asociaciones del Ministerio del Interior)

Variables	Descripción	Tipo de variable
Nombre	Cuál es el nombre de la ONG	Texto
Sectores	Para qué sectores está cualificada	Categórico
CIF	Cuál es el número de identificación de cliente de la ONG	Texto

Fuente: https://www.aecid.es/EN/aecid/our-partners/ngdo/accreditation

El conjunto de datos proporciona dos ID posibles para enlazar: el nombre de la organización y su número de identificación fiscal (NIF). Ambos se pueden utilizar para vincular los datos a los datos de subvenciones de la IGAE. Los datos constan de tres variables, dos de las cuales son identificaciones y una especifica los sectores precisos en los que la ONG está cualificada para operar. Partiendo de esta información, se pueden crear dos variables binarias: 1) Si la ONG ha sido revisada y 2) si la ONG está actuando en la misma área para la que estaba cualificada (por ejemplo, la ONG estaba cualificada para el sector sanitario, pero recibe subvenciones para el sector de educación). Debido al bajo número de entidades, es poco probable que se produzcan cambios significativos en las calificaciones de riesgo previstas. Sin embargo, si el conjunto de datos principal de la BDNS se filtra solo para las ONG, esta información podría influir en los resultados para este sector.

Evaluaciones de ONG

El tercer conjunto de datos que merece tenerse en cuenta es el de la Fundación Lealtad. Se trata de un evaluador independiente de ONG, que analiza la gestión, gobernanza, uso de fondos, situación económica, voluntariado y transparencia de las ONG. En el sitio web de la fundación existe un archivo PDF descargable con la lista de todas las ONG evaluadas positivamente. Sin embargo, esta lista contiene información limitada más allá del nombre de las organizaciones. Por tanto, un enfoque más eficaz sería acceder a las páginas HTML de cada organización y analizar los datos manualmente. Existe la posibilidad de analizar información de archivos PDF estandarizados denominados «informes completos» para cada ONG. El directorio de variables se detalla en la Tabla 3.4.

Tabla 3.4. Directorio de variables (Fundación Lealtad)

Variables	Descripción	Tipo de variable
Nombre	El nombre de las ONG	Texto
Sectores	Sectores en los que opera	Categórico
NIF	El código NIF de la ONG	Texto
Ingresos	Los ingresos anuales de la organización + fuentes	Numérico + categórico
Gastos	Los gastos anuales de la organización + tipos de gastos	Numérico + categórico
Año	Año de origen de la organización	Fecha
Beneficiarios	El número total y el tipo de beneficiarios de esta ONG.	Numérico
Socios	Número de socios que tiene la ONG	Numérico
Empleados	Número de empleados que tiene la ONG	Numérico
Voluntarios	Número de voluntarios que tiene la ONG	Numérico
NIF	El número de NIF de la organización	Texto
Puestos de gerencia	Persona/s que representan a la gerencia de esta ONG	Texto
Contactos	Correo electrónico, teléfono, dirección de la organización	Texto
Zona geográfica	Dónde opera la ONG	Texto

Fuente: https://www.fundacionlealtad.org/ong/a-toda-vela/

Las principales ID mediante las cuales las organizaciones pueden vincularse a los conjuntos de datos de la IGAE son el nombre de la organización y el NIF. Si bien el nombre está disponible en archivos HTML y PDF, el NIF se almacena en el PDF del informe completo. Los datos sobre ingresos, gastos, sector de actividad, año de origen, así como el número de beneficiarios, socios y empleados pueden agregarse a la información de antecedentes para el análisis. Como antes, se puede crear una variable binaria que refleje si la organización en cuestión está verificada o no por la Fundación Lealtad. Además de la información general de antecedentes, se pueden extraer algunos indicadores adicionales de este conjunto de datos. Por ejemplo, se debe tener en cuenta la proporción de gastos para evaluar cuánto se gasta la ONG en su propia gestión en comparación con su misión. Un gasto elevado en gestión podría ser una señal de calificación de riesgo más alta aunque, por sí solo, no sería un indicador de fraude o irregularidades. Las personas en cargos directivos, cuando se cruzan con otros conjuntos de datos (por ejemplo, personas políticamente expuestas), pueden proporcionar información sobre posibles conflictos de intereses.

Cruzar datos personales para rastrear conexiones y conflictos de interés

El segundo grupo de conjuntos de datos que podrían mejorar el modelo de riesgo de la IGAE, descrito en el Capítulo 2, son los datos sobre conexiones personales y conflictos de interés. Vincular datos sobre conexiones personales entre los sectores público y privado abre la posibilidad de rastrear conflictos de interés. Estos datos se pueden analizar mediante análisis de redes, para identificar si existen conexiones entre personas políticamente expuestas y propietarios de las empresas que reciben subvenciones y ayudas. Ya se trataron en el grupo anterior algunas fuentes posibles. Los siguientes apartados se centrarán en el Registro de Propiedad efectiva y FICESA, la base de datos de altos cargos y directivos españoles.

Registro de Propiedad Efectiva (BO)

El registro de BO proporciona información de más de 5 000 000 de organizaciones registradas desde 2009. La lista corta de variables se proporciona en la Tabla 3.2. No hay un conjunto de datos completo de dominio público, pero la fuente – una plataforma online para consultar y analizar el Boletín Oficial del Registro Mercantil (LibreBOR) - proporciona una API y un *script* de Python para analizar los datos.[3] Es

posible seleccionar aquellas organizaciones que aparecen en los conjuntos de datos de la IGAE, sin analizar todo el conjunto de datos, lo que hará más eficiente el tiempo de tratamiento.

Tabla 3.5. Directorio de variables en el registro BO

Variables	Descripción	Tipo de variable
Denominación actual y anterior	El nombre de la empresa, cuáles son los nombres anteriores	Texto
Domicilio social	La oficina oficial está registrada	Texto
Forma jurídica	La forma jurídica de la empresa	Categórico
Provincia	Provincia en la que opera la empresa	Texto
Puestos directivos	Nombres de la/s persona/s en puestos directivos	Texto
Fecha de disolución y motivo	Si la empresa se cerró o se desintegró; cuándo y por qué sucedió	Fecha + texto
Datos de registro	Información adicional sobre el registro de la empresa	Texto
Enlaces a las fuentes oficiales	Fuente oficial de la que proceden los datos	Texto
Propietarios efectivos	Lista de nombres de los propietarios finales de la empresa.	Texto

Fuente: https://docs.librebor.me/

La IGAE tiene dos formas de cruzar los conjuntos de datos de la BDNS con el registro BO: 1) Por nombre de la organización, o 2) por NIF del beneficiario. Como alternativa, es posible agregar datos por provincia y enlazar estos números agregados (por ejemplo, tamaño medio de la empresa) por ubicación particular. El conjunto de datos del BO contiene mucha información de antecedentes para organizaciones, pero la más relevante son los puestos directivos, las organizaciones asociadas y los propietarios efectivos finales. Los datos de propiedad se utilizan mejor cuando se comparan con otros conjuntos de datos, en particular, listas de titulares de cargos políticos (consulte la siguiente sección).

Además, la IGAE puede utilizar parte de la información de antecedentes como predictores de riesgo en sí mismos. Cuando los nombres de los propietarios efectivos de entidades beneficiarias de subvenciones se cruzan con los de los titulares de cargos públicos, es posible identificar conflictos de interés directos (es decir, cuando el beneficiario trabaja para el organismo que concede la subvención) o formas indirectas de conflicto potencial (es decir, cuando el titular del cargo político relacionado trabaja en una organización de nivel superior o en un órgano de supervisión de la organización concedente). Cuando se analizan los datos de propiedad por sí solos, la información sobre las empresas asociadas con el beneficiario puede revelar riesgos si se cruza con otros conjuntos de datos (por ejemplo, formas complejas de conflictos de interés y factores de riesgo relacionados). [4]

Base de datos de altos cargos

La siguiente fuente es una base de datos de altos cargos y directivos de España llamada FICESA. Esta fuente contiene datos relacionados con altos cargos públicos en una amplia gama de organizaciones públicas: Secretarías de Estado, Subsecretarías, Direcciones Generales y Subdirecciones, Oficinas de Presupuestos, así como diferentes órganos judiciales a escala nacional, regional y local. No son datos de dominio público y los datos deben solicitarse al titular de los datos rellenando un formulario. Por tanto, el formato de los datos y las variables que contiene el conjunto de datos no está claro. No hubo respuesta a los intentos de contactar con la fuente. Se supone que la IGAE podría obtener acceso a la base de datos completa como una descarga masiva.

La única identificación por la que se puede vincular este conjunto de datos son los nombres y, si están disponibles, datos personales adicionales, como la fecha de nacimiento. Si el conjunto de datos de la BDNS contuviera datos sobre propietarios efectivos, como se indica anteriormente, los datos sobre cargos oficiales podrían cruzarse por nombres de personas. Vincular los conjuntos de datos de la IGAE a la información sobre los titulares de cargos de alto nivel crea la posibilidad de realizar análisis de red y ver

si existen conflictos de interés entre las organizaciones privadas que reciben subvenciones y los organismos públicos que las conceden. Es especialmente útil utilizar el registro BO para buscar todas las organizaciones asociadas y analizar si están conectadas con personas políticamente expuestas. Por ejemplo, que aunque la organización que recibe la subvención no esté relacionada con nadie de organismos oficiales, una de sus organizaciones vinculadas sí podría estarlo.

Cruzar datos sobre fiabilidad organizativa e infracciones para recopilar riesgos en diferentes dominios

Los conjuntos de datos con información sobre fiabilidad organizativa e infracciones de normas o leyes es el tercer grupo de datos que podría ayudar a la IGAE a fortalecer su modelo de riesgo para evaluar los riesgos de fraude de subvenciones. Este grupo quedó cubierto parcialmente en la sección sobre datos de ONG acreditadas. Además, en este grupo, hay conjuntos de datos sobre suspensiones de pagos e impuestos. Cruzar los datos sobre la fiabilidad organizativa y la infracción de normas arroja luz sobre nuevas dimensiones de riesgos de fraude relacionados con otros dominios. Estos conjuntos de datos pueden ayudar a predecir los riesgos de fraude en las subvenciones al explotar las correlaciones entre la fiabilidad de las organizaciones acreditadas, los comportamientos de cumplimiento de normas (deudas fiscales, suspensiones de pagos, etc.) y el fraude en las subvenciones. Partiendo de propuestas anteriores, la siguiente sección se centra en el Registro Público Concursal, los datos fiscales de la AEAT y los datos contables de CINCOnet.

Registro Público Concursal

El primer conjunto de datos de este grupo, calificado previamente como una prioridad media para la IGAE, es el Registro Público Concursal. La fuente incluye información de resoluciones procesales, suspensiones de pagos y acuerdos extrajudiciales. Los datos HTML se pueden analizar después de filtrar por provincia o tribunal. Desgraciadamente, por motivos desconocidos, el filtrado no funciona correctamente en el sitio, lo que genera errores en la página. Aun así, el directorio aproximado de variables se presenta en la Tabla 3.6.

Tabla 3.6. Directorio de variables (Registro Público Concursal)

Variables	Descripción	Tipo de variable
Nombre	El nombre de la empresa	Texto
Documento identificativo	La ID del documento concursal	Texto
Deudor	Si la empresa es una deuda o no	Binario
Incapacitada	Si la empresa está incapacitada o no	Binario
Administrador	Si la empresa es administradora de la quiebra o no	Binario

Este conjunto de datos se puede cruzar con los datos de subvenciones de la IGAE por el nombre de la organización o por código NIF. La fuente no brinda la oportunidad de revisar todos los casos, lo que requiere un filtrado de antemano, por lo que la forma más fácil de establecer un filtro es por provincia. La información más relevante para las evaluaciones del riesgo de fraude son los detalles sobre la suspensión de pagos. La fuente proporciona ubicación, nombre de la organización, tribunal, juez y NIF u otros identificadores de las empresas. Lamentablemente, no hay información sobre la fecha de los procedimientos concursales, lo que sería especialmente importante para analizar las subvenciones y ayudas anteriores. Después de lcruzar los datos, el indicador de riesgo más relevante para la IGAE sería la variable binaria («bandera») que refleja si el concesionario estaba o se encuentra actualmente en estado concursal. Dicha información sobre la situación de una empresa podría indicar que el beneficiario

hará un mal uso de la subvención o ayuda concedida, o al menos que se gestionará esta de forma inadecuada debido a otras presiones organizativas.

Datos fiscales

El segundo conjunto de datos sobre incumplimiento de normas son los datos de la Agencia Estatal de Administración Tributaria (AEAT). Este es un conjunto de datos con acceso restringido, y solo las estadísticas agregadas son de dominio público. Una vez más, para lo que se trata a continuación, se supuso que la IGAE puede obtener acceso completo a la base de datos, para incorporar dichos datos en su modelo de riesgo. Según las notas publicadas por la AEAT, se dispone de datos en formato desagregado que pueden ser facilitados previa solicitud. Los datos agregados cubren la presentación de declaraciones fiscales, pago de impuestos, deudas y tasas, certificados de impuestos, declaraciones fiscales, etc.

Debido al acceso restringido a los conjuntos de datos, no está claro si las ID son las mismas que en el conjunto de datos BDNS, pero lo más probable es que las entidades se puedan enlazarse por nombre o por NIF del beneficiario. La información sobre el pago puntual de impuestos, deudas y otros cargos es la más relevante para enriquecer los modelos predictivos sobre riesgos de fraude. Los retrasos en el pago de impuestos, así como la existencia de deudas en una determinada empresa (o asociadas) podría ser una señal de mayores riesgos.

Información contable

El tercer conjunto de datos que pertenece a este grupo son los datos contables y presupuestarios de CINCOnet, considerados de alta prioridad para la IGAE, y las mejoras en el modelo de riesgo. Los datos incluyen operaciones de gasto e importe total de gastos en el año en curso, importe de ingresos en el año en curso, liquidez, operaciones no presupuestarias, gastos de terceros, datos generales de terceros, etc. Como los datos de la AEAT, estos datos no son de dominio público; sin embargo, el Ministerio de Hacienda y Función Pública administra CINCO.net y la IGAE tiene acceso directo.

Las entidades de esta base de datos se pueden cruzar por nombre o NIF del beneficiario con la BDNS. Sin embargo, debido al acceso restringido a los datos, es difícil evaluar la calidad y el contenido de las variables. Además de la información general sobre ingresos y gastos, CINCOnet proporciona datos sobre el reintegro de otras subvenciones concedidas por diferentes organizaciones en España. Esto puede ser especialmente útil en la evaluación de riesgos potenciales en la provisión de subvenciones y ayudas en el futuro, como la doble financiación de operaciones o el gran valor de subvenciones recibidas comparado con los ingresos.

Cruzar datos de contratos públicos y otras subvenciones permite rastrear la doble financiación y los riesgos asociados

El grupo final de conjuntos de datos abarca un elenco diverso de datos sobre contratos públicos y otras subvenciones y financiación. Ligar datos de otros fondos y contratos permitiría a la IGAE hacer una referencia cruzada del gasto y desarrollar dimensiones de riesgo adicionales. Por ejemplo, puede ayudar a identificar subvenciones acumuladas para las mismas actividades, que deben considerarse un factor de riesgo. Los contratos públicos recibidos por una empresa pueden puntuarse utilizando indicadores de riesgo de corrupción y luego relacionados con riesgos de subvenciones. Por ejemplo, una empresa o agencia (tercero, concedente, concesionario) que participa en licitaciones de alto riesgo también puede presentar riesgo cuando se trata de subvenciones. Este grupo incluye grupos de datos de la Asociación Española de Fundaciones (AEF), Fondos de la Unión Europea y datos de contratación pública.

Datos de fundaciones

Los datos de la AEF proporcionan información sobre las fundaciones que conceden subvenciones, entre ella: Su tipo de actividad, zonas geográficas, tipo de beneficiarios, fecha de creación y órganos de gestión. El directorio de variables se presenta en la Tabla 3.7. Los datos son de acceso abierto y se pueden descargar fácilmente en formato Excel o PDF. En total hay 15 840 fundaciones recogidas en el directorio.

Tabla 3.7. Directorio de variables de la Asociación Española de Fundaciones (AEF)

Variables	Descripción	Tipo de variable
Nombre	Cuál es el nombre de la fundación	Texto
Protectorado	Al amparo de qué ministerio/agencia protectorado se encuentra esta fundación	Texto
Año	Año de creación	Fecha
Contactos	Cuáles son los datos de contacto de la fundación (correo electrónico, teléfono)	Texto
Dirección	Dónde opera la fundación	Texto

Fuente: http://www.fundaciones.es/es/buscador-fundaciones

Relacionar este conjunto de datos con la BDNS implica varios pasos. Primero, todas las observaciones deben filtrarse por tipo de beneficiario, utilizando el filtrado en tiempo real, ya que el tipo de beneficiario no es un campo de datos en el archivo descargable. En segundo lugar, la ubicación particular debe coincidir con la ubicación de los concedentes o beneficiarios. Esto no proporcionará la información exacta sobre si el beneficiario recibió otra subvención de una fundación determinada, pero indica la presencia de la fundación en el mismo lugar con los mismos tipos de beneficiarios.

La información más relevante para que la IGAE evalúe los riesgos sería si alguno de los beneficiarios recibió doble financiación para las mismas actividades. Para rastrear con precisión dichos riesgos, es necesario verificar los beneficiarios exactamente por sus identificaciones. Sin embargo, esta fuente no proporciona información tan detallada. Por tanto, solo la información agregada, que es mucho más imprecisa, se puede utilizar desde esta fuente. La presencia de una fundación que apoye actividades similares en la misma localidad (provincia) que el concedente o concesionario aumenta la probabilidad de recibir doble financiación.

Datos de fondos de la Unión Europea (UE)

El siguiente conjunto de datos relevante para que la IGAE valore su conexión con los datos de la BDNS, con prioridad media, sin datos para Fondos de la Unión Europea. El gobierno español y la Comisión Europea proporcionan los datos y disponen de registros desde 2007 a 2020. Los datos son de fácil acceso y se pueden descargar en formato Excel. El directorio de variables relevantes se presenta en la Tabla 3.8.

Tabla 3.8. Directorio de variables (ayudas de la Unión Europea)

Variables	Descripción	Tipo de variable
Referencias presupuestarias	La ID de referencia del presupuesto para esta subvención	Texto
Objeto de subvención o contrato	El fin/objeto de esta subvención	Texto
Nombre del beneficiario	El nombre del beneficiario	Texto
Número de IVA (NIF)	El número de IVA del beneficiario (NIF)	Texto
Importe contratado	El importe que se contrató al beneficiario	Numérico
Número de compromisos presupuestarios	El número de compromisos presupuestarios que tiene el beneficiario	Numérico
Nombre del programa	El nombre del programa bajo el cual se asignó la subvención	Texto
Departamento responsable	El departamento responsable de la asignación de subvenciones	Texto
Fecha de inicio y finalización del proyecto	La fecha de inicio y finalización del proyecto	Fecha

Fuente: https://ec.europa.eu/budget/financial-transparency-system/analysis.html

Los datos proporcionan un código de IVA como identificación para las organizaciones, que se puede convertir en un NIF eliminando las dos primeras letras. Alternativamente, se pueden usar los nombres de organizaciones para enlazar. El número de compromisos presupuestarios, objeto de subvenciones o contratos, así como las fechas de inicio y finalización del proyecto son especialmente relevantes para identificar si el beneficiario recibió financiación de la UE para el mismo proyecto que su subvención española. La doble financiación es una práctica fraudulenta cuando el mismo proyecto es financiado más de una vez por diferentes subvencionadores, sin proporcionar información a ambos sobre las aportaciones realizadas por el otro. Por tanto, el proyecto podría implantarse, pero el dinero público adicional desembolsado no se utiliza como se esperaba.

Datos de contratación pública

La última fuente de datos que la IGAE podría valorar vincular con sus conjuntos de datos son los datos de contratación pública nacional. El portal opentender.eu contiene estos datos recopilados de dos fuentes gubernamentales oficiales (el Ministerio de Hacienda y Función Pública y la Plataforma de Contratación), así como el Tender Electronic Daily (TED). Los datos contienen toda la información disponible públicamente sobre licitaciones, contratos, licitadores, contratantes y contratistas necesaria para calcular el indicador de riesgo de corrupción (ver Recuadro 3.1). El directorio de variables relevantes se presenta en la Tabla 3.9.

Tabla 3.9. Directorio de variables (datos de contratación pública)

Variables	Descripción	Tipo de variable
ID del contratista	ID única del proveedor	Texto
ID del contratante	ID única del comprador	Texto
Nombre del contratista	Nombre del proveedor que gana el contrato	Texto
Nombre del contratante	Nombre del comprador que convoca la licitación	Texto
Numero de ofertas	Cuántas ofertas se hicieron por licitación	Numérico
Tipo de procedimiento	¿El tipo de procedimiento es abierto o restringido?	Categórico
Convocatoria pública	¿Estaba la licitación a disposición del público?	Categórico
Duración de la presentación de la oferta	La duración entre la fecha de inicio y finalización de presentación de la oferta	Numérico
Duración del período de decisión	La duración entre la fecha de finalización de presentación de la oferta y la decisión	Numérico
Conexiones	¿Existen conexiones registradas entre el contratista y la autoridad de contrataciones?	Categórico

Fuente: Platforma de Contratacion https://contrataciondelestado.es/; Portal Institucional Del Ministerio De Hacienda y Funcion Pública: https://www.hacienda.gob.es; Tenders electronic daily: http://ted.europa.eu.

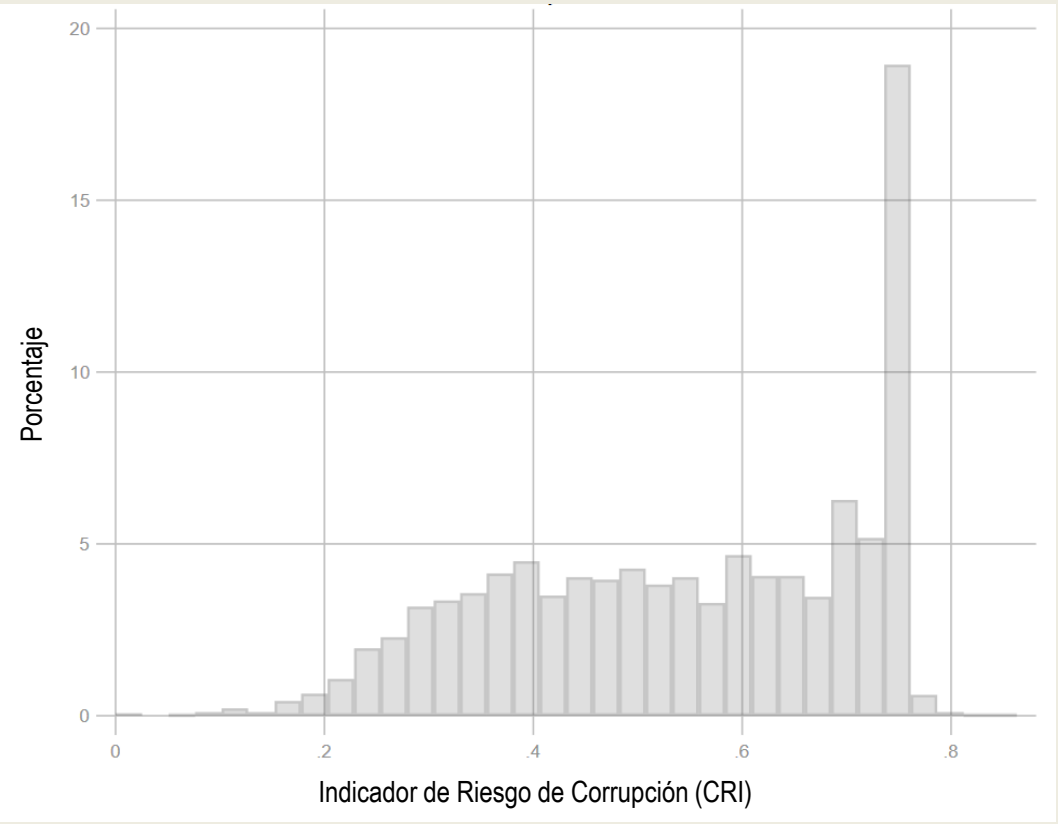
Las ID de contratistas son las mismas que los NIF de los beneficiarios. Por tanto, esta ID se puede utilizar para cruzar datos. De forma alternativa, los nombres de organizaciones, así como los nombres de concedentes, pueden cruzarse con los contratantes o proveedores del conjunto de datos de contrataciones. Para evaluar si los contratos ganados por empresas licitadoras, o las licitaciones efectuadas por contratantes públicos son proclives a la corrupción, se pueden usar indicadores de aproximación (proxies) de corrupción. Por ejemplo, la licitación única en mercados competitivos, el tipo de procedimiento utilizado, la publicidad de la convocatoria de licitaciones, la duración del anuncio de licitación y el período de decisión, así como conexiones entre el contratista y la autoridad de contrataciones. La recopilación de riesgos de corrupción en contrataciones públicas en las actividades de contrataciones de los beneficiarios o concedentes puede arrojar más luz sobre los riesgos de fraude en subvenciones, ya que es verosímil que las organizaciones con riesgo en un dominio también lo sean en un otros dominio relacionado. Esta lógica de análisis se demuestra empíricamente en Recuadro 3.1.

Recuadro 3.1. Cruzar datos de subvenciones de la IGAE con datos de contratación pública (conjunto de datos de opentender.eu)

El indicador de riesgo de corrupción (CRI) muestra la restricción deliberada de la competencia en licitaciones de contratación pública en beneficio de una empresa licitadora vinculada. La metodología de CRI utiliza datos administrativos para calcular las puntuaciones de riesgo de corrupción para cada contrato. Basándose en la metodología desarrollada por (Fazekas y Kocsis 2017), el criterio de selección de indicadores de riesgo de contrataciones es el grado de asociación con una restricción injustificada de la competencia, es decir, licitación única en mercados competitivos. Incluye varios indicadores próximos (*proxies*) de corrupción además de la licitación única, como el riesgo del tipo de procedimiento cerrado de contrataciones, la falta de publicidad de las licitaciones, el registro de residencia en paraísos fiscales de los proveedores, la dependencia de la autoridad de contrataciones del proveedor (es decir, la captura por el agente) y la duración del anuncio de licitación y los períodos de decisión.

Utilizamos la identificación fiscal (NIF) de los proveedores para ligar el conjunto de datos de subvenciones con el conjunto de datos de contrataciones públicas una vez limpias. Después de limpiar NIFs de entradas sin sentido, las calificaciones de riesgo de fraude de subvenciones se agregaron para cada proveedor y se vincularon directamente con el conjunto de datos de contrataciones. Hubo 103 872 contratos adjudicados a 6 408 contratistas que habían recibido una subvención. La Figura 3.1 muestra la distribución CRI agregada para contratistas subvencionados, sin incluir los contratistas con menos de 3 contratos. Hay una calificación media CRI de 0,55, considerablemente más alta que la media nacional.

Figura 3.1. Distribución CRI (contratistas)



Cruzar el conjunto de datos de subvenciones con el conjunto de datos de contrataciones públicas permite obtener conocimientos más profundos sobre las relaciones entre las calificaciones de riesgo. Se han efectuado análisis de regresión lineal y no lineal, incluidos los controles de ubicación del contratante, tipo de contratante, tipo de mercado (sectores CPV), tipo de contrato y año de licitación. Ambos modelos en la Tabla 3.10 muestran una correlación positiva entre las calificaciones de riesgo de corrupción en las contrataciones y los riesgos de fraude en subvenciones. Sin embargo, el modelo 2 parece ajustarse mejor al capturar la no linealidad de esta relación. En la Figura 3.2 mostramos los márgenes predictivos de modelar el CRI en una relación cuadrática con el Riesgo de Fraude de Subvenciones. Estos resultados de regresión simple nos aseguran la validez de ambas calificaciones de riesgo, ya que están alineadas y transmiten un mensaje similar: que las calificaciones de riesgo de corrupción más altas se correlacionan positivamente con riesgos de fraude de subvenciones más altos. Además, la asociación es especialmente fuerte cuando los riesgos de corrupción en contrataciones públicas están por encima de la media de la muestra.

Figura 3.2. Correlación entre CRI y Riesgos de Fraude de Subvenciones (márgenes predictivos)

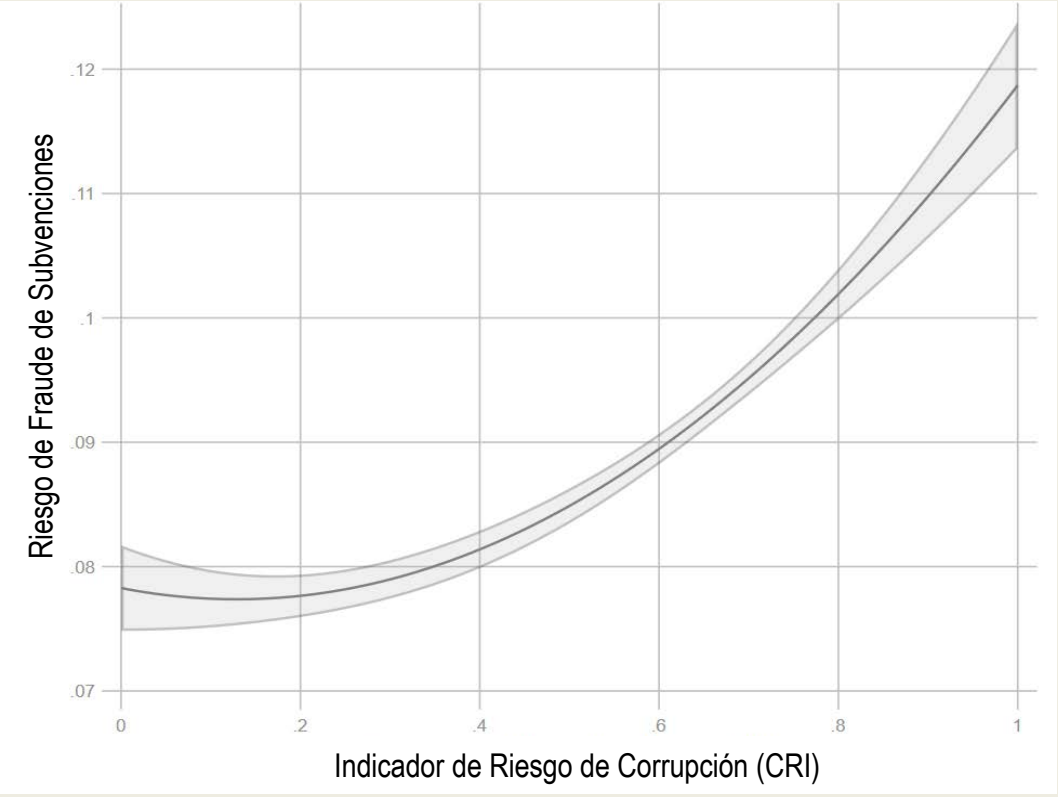


Tabla 3.10. Correlación entre CRI y Riesgo de Fraude de Subvenciones

Variable dependiente	Riesgo de Fraude de Subvenciones	
Modelo	**(1)**	**(2)**
Muestra	**Concedido**	**Concedido**
CRI	0,036*** (0,002)	-0,014 (0,021)
CRI^2		0,054** (0,024)
Controles	✓	✓
Observaciones	103 151	103 151
R^2	0,1719	0,1721

Notas: La regresión incluye controles para valores de contrato, tipo de contrato, tipo de comprador, ubicación del comprador, mercado, tipo de contrato y año de licitación. Errores estándar robustos entre paréntesis *** p<0,01, ** p<0,05, * p<0,1.

Fuente: Fazekas, M. y G. Kocsis (2017[1]), «Revelar la corrupción de alto nivel: indicadores de riesgo de corrupción de objetivos transnacionales que utilizan datos de contratación pública». British Journal of Political Science 50 (1): 155–64, http://dx.doi.org/10.1017/s0007123417000461

Ventajas de utilizar múltiples conjuntos de datos

Este capítulo ofrece una descripción detallada de cómo y por qué diferentes conjuntos de datos se pueden vincular a los conjuntos de datos actualmente existentes de la IGAE, con especial atención a los prometedores indicadores de riesgo de fraude habilitados por los nuevos datos. Estos nuevos indicadores detectan principalmente el comportamiento de los actores, en lugar de simples características de antecedentes, lo que permite una evaluación de riesgos mucho más precisa. Sin embargo, el cruce de datos no solo permite calcular nuevos indicadores en una base de datos y vincularlos entre sí, sino también crear nuevos indicadores basándose en múltiples conjuntos de datos. Estos indicadores complejos ofrecen información adicional sobre dimensiones relevantes de riesgo. También representan una medida más robusta del comportamiento del actor, porque varias fuentes que apuntan al mismo comportamiento tienen mayor validez que un solo conjunto de datos.

El uso de múltiples conjuntos de datos es crucial para caracterizar de manera integral comportamientos complejos de fraude, así como para reducir el índice de falsos positivos, que son frecuentes en modelos simples (Fazekas, M., Ugale, G, & Zhao, A., 2019[2]). Combinar varios indicadores derivados de diferentes conjuntos de datos se considera una buena práctica en la medición del riesgo, ya que permite la triangulación de la medición. En otras palabras, permite aumentar la convergencia de validación. Los falsos positivos son omnipresentes en las evaluaciones de riesgo simples, ya que muchos indicadores simplemente apuntan a posibles irregularidades en lugar de fraudes reales. Además, los indicadores de conflicto de interés generalmente utilizados suelen indicar la presencia de un conflicto potencial en lugar de un conflicto real que represente el abuso de una situación para un beneficio personal indebido. Sin embargo, cuando la información sobre conflictos de interés se combina con datos sobre resultados, como acumulación de subvenciones o desempeño financiero anómalo, la combinación de indicadores proporciona una mayor validez al enfoque de medición.

Cruzar conjuntos de datos que representan múltiples dimensiones de relaciones también puede impulsar el uso de análisis avanzados de red de múltiples capas. Estas relaciones de varios niveles pueden abarcar conexiones entre empresas privadas y organizaciones públicas que otorgan subvenciones a través de una variedad de relaciones contractuales, o vínculos entre los propietarios reales de empresas y personas políticamente expuestas que tienen cargos del sector público. Varias conexiones de red establecidas mediante el uso de conjuntos de datos cruzados de gestión a gran escala también permiten realizar un seguimiento de los cambios temporales en las conexiones entre entidades e individuos potencialmente de riesgo, lo que aumenta la sofisticación analítica del modelado de riesgos.

Conclusión

Esta sección ha revisado una amplia variedad de conjuntos de datos adicionales útiles para el conjunto de datos actual de la IGAE. Al hacerlo, estableció una hoja de ruta para la captura de datos y el cruce que optimiza el valor analítico para la IGAE. De los conjuntos de datos revisados, la información empresarial sobre registro, propiedad y finanzas representa el mayor potencial para perfeccionar aún más el modelo de evaluación de riesgo de fraude. Estos conjuntos de datos se pueden cruzar fácilmente con datos internos de la IGAE utilizando las ID de registros empresariales. Además, cruzar datos de contratación pública con datos de subvenciones, también demostrado mediante el análisis de conjuntos de datos fácilmente disponibles, puede añadir un gran valor, ya que 2 conjuntos de factores de riesgo se pueden triangular entre sí para producir una evaluación de riesgo más fiable. Una vez que estos conjuntos de datos de prioridad alta se incorporan a la canalización de datos de la IGAE, también se pueden considerar otros conjuntos de datos, como el registro concursal.

Referencias

Fazekas, M., Ugale, G, & Zhao, A. (2019), *Analytics or Integrity: Data-Driven Decisions for Enhancing Corruption and Fraud Risk Assessments*, OECD Publishing, Paris, https://www.oecd.org/gov/ethics/analytics-for-integrity.pdf. [2]

Fazekas, M. and G. Kocsis (2017), "Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data", *British Journal of Political Science*, Vol. 50/1, pp. 155-164, http://dx.doi.org/10.1017/s0007123417000461. [1]

Notas

[1] En algunos casos, se presume que determinada información está presente en los conjuntos de datos de la IGAE. Sin embargo, la confirmación de esto no fue posible debido a la anonimización de la mayoría de las bases de datos.

[2] El acceso al conjunto de datos está restringido y exige pagar una tarifa por cada organización y recibir un certificado digital. Solo se permite el acceso libre a los datos agregados por sector, año o sector empresarial. La única información disponible a escala empresarial sin restricciones adicionales es el estado de la empresa (operativa o no). Para que la IGAE utilice estos datos, necesitaría obtener acceso total al conjunto de datos completo y actual, ya sea pagando la tarifa de acceso masivo o llegando a un acuerdo especial con el proveedor de datos del gobierno. También existen alternativas públicas de fácil acceso, por ejemplo, opencorporates.com, que es una empresa social privada que tiene como objetivo hacer que todos los datos empresariales sean fácilmente accesibles en todo el mundo.

[3] Consulte https://docs.librebor.me/python/.

[4] Debido al acceso restringido a la fuente, no está claro si la información sobre los propietarios efectivos está allí. Sin embargo, está presente en el registro mercantil, por lo que es razonable esperar que también contenga una variable en *LibreBOR*. En caso de que no lo sea, la información se puede obtener del registro de la empresa después de recibir un certificado electrónico.

Anexo A. Estadísticas descriptivas de variables en el conjunto de datos limpio

La tabla muestra el mínimo, la media, la mediana y el máximo para variables numéricas y el número de casos por las categorías más frecuentes para variables categóricas.

ABIERTO_CON420 S : 45701 N : 1004769	AUDAESTADO_CON490 N : 728696 S : 321774	FINALIDAD_CON540 12 : 408525 6 : 120412 16 : 84776 5 : 83064 11 : 81101 18 : 60650 (Otro) : 211942	NOMINATIVA_CON610 N : 659757 S : 390713
PUBLICABLE_CON620 0 : 18819 1 : 1031651	IMPACTOGENERO_CON630 1 : 54816 2 : 836210 3 : 159301 4 : 143	FECHA_ACTUALIZACION Mín. : 31.03.2000 Media : 12.04.2019 Mediana : 05.03.2020 Máx. : 02.06.2021	PAIS_TER100 ES :1048239 AR : 259 VE : 144 FR : 141 IE : 114 DE : 104 (Otro) : 1469
ID_TER110 Longitud: 1050470 Clase: carácter	PAISDOM_TER250 ES : 1048238 AR : 261 VE : 146 FR : 137 IE : 114 DE : 104 (Otro) : 1470	NATURALEZA_TER280 SOCIEDADES DE RESPONSABILIDA LIMITADA : 293159 ASOCIACIONES : 186563 CORPORACIONES LOCALES : 166903 SOCIEDADES CIVILES : 88221 (Otro) : 315624	TIPOBEN_TER290 FSA : 5300 GRA : 6395 JSA : 512945 PFA : 435385
REGION_TER310 ES :179660 ES41 : 83013 ES51 : 65044 ES43 : 55058 ES511 : 44993 ES425 : 39884 (Otro): 582818	ID_COS Longitud: 1050470 Clase: carácter	DAT_COS_CSU210 Mín. : 01.01.2018 Máx. : 31.12.2020 Mediana : 28.12.2018 Media : 26.06.2019	COSTE_ACT_CSU240 Mín. : 0.000e+00 Máx. : 3.120e+10 Mediana : 3.000e+03 Media : 6.343e+04
IMPORTE_CONCE_CSU220 Mín. : 0 Máx. : 139112532 Mediana : 2296 Media : 23886	AYUDA_EQUI_CSU250 Mín. : 0 Máx. : 139112532 Mediana : 2296 Media : 23886	REGION_CSU260 ES : 179702 ES41 : 83266 ES51 : 69608 ES43 : 65568 ES511 : 45443 ES425 : 39871 (Otro) : 567012	Año Mín. : 2018 Mediana : 2018 Máx. : 2020
FECHAPAGO_PAG210 Mín. : 25.07.2017 Mediana : 11.07.2019	IMPORTE_PAG220 Mín. : 0 Mediana : 2000	RETENCION_PAG230 N : 1048245 S : 2225	CON550 A : 179727 S : 77071

Media : 03.10.2019 Máx. : 01.06.2021	Media : 15897 Máx. : 105745000		J : 67460 (Otro) : 688175
CON560 OTROS, SUBV : 200 PREST, SUBV : 1982 SUBV : 1045030 SUBV, OTROS : 109 SUBV, PREST : 3058 SUBV, VENTA : 91	CON580 PFA : 501971 JSA : 398807 FSA : 16830 PFA, GRA : 15287 GRA, PFA : 13359 JSA, PFA : 12484 (Otro) : 91732	CON570 ES : 204438 ES42 : 135555 ES41 : 90127 ES51 : 67554 ES43 : 65829 ES61 : 39539 (Otro) : 447428	SAN_dum 0 : 1049439 1 : 1031
Month_CSU210 12 :215890 9 :156651 6 :112147 10 :109546 11 : 98538 7 : 87721 (Otro): 269977	Nawards_TER_110 Mín. : 1,00 Mediana : 9,00 Media : 17,88 Máx. :193,00	Amount_awards_TER_110 Mín. : 1 Mediana : 95929 Media : 85060 Máx. :139915	NATIONAL_CSU260 0:870533 1:179937
REGIONAL_CSU260 0:1013951 1: 36519	MUNICIPAL_CSU260 0:216456 1:834014	NATIONAL_TER310 0:870551 1:179919	REGIONAL_TER310 0:1014840 1: 35630
MUNICIPAL_TER310 0:209850 1:840620	LOCAL_IMPL 0:100402 1:950068	SECTOR_CON550_AGR 0:589338 1:461132	SECTOR_CON550_MINING 0:995418 1: 55052
SECTOR_CON550_MANUF 0:968989 1: 81481	SECTOR_CON550_ELECTR 0:995312 1: 55158	SECTOR_CON550_WATER 0:991202 1: 59268	SECTOR_CON550_CONSTR 0:975388 1: 75082
SECTOR_CON550_RETAIL 0:971928 1: 78542	SECTOR_CON550_TRANSP 0:982215 1: 68255	SECTOR_CON550_ACCOM 0:965384 1: 85086	SECTOR_CON550_INFO 0:914043 1:136427
SECTOR_CON550_FIN 0:992458 1: 58012	SECTOR_CON550_RESTAT 0:996060 1: 54410	SECTOR_CON550_SCI 0:963038 1: 87432	SECTOR_CON550_ADMIN 0:946261 1:104209
SECTOR_CON550_SECUR 0:972335 1: 78135	SECTOR_CON550_EDUC 0:943212 1:107258	SECTOR_CON550_HEALTH 0:925630 1:124840	SECTOR_CON550_ART 0:887795 1:162675
SECTOR_CON550_OTHER 0:866250 1:184220	SECTOR_CON550_HOUSEHOL D 0:1029872 1: 20598	SECTOR_CON550_EXTRATER 0:1018305 1: 32165	

Anexo B. Directorio completo de variables en el conjunto de datos sin limpiar

Variable	Descripción breve	Descripción adicional de la variable, si procede	Tipo
ADMINISTRACION_ANTE	Unidad de administración	Qué unidad de administración proporciona la convocatoria	carácter
DEPARTAMENTO_ANT	Departamento	Qué departamento proporciona la convocatoria	carácter
ORGANO_ANTE	Órgano	Qué órgano oficial proporciona la convocatoria	carácter
ADMINISTRACION	Unidad de administración	Qué unidad de administración proporciona la convocatoria	carácter
DEPARTAMENTO	Departamento	Qué departamento proporciona la convocatoria	carácter
ORGANO	Órgano	Qué órgano oficial proporciona la convocatoria	carácter
DIR3_CON710	Entidad concedente	Identificación de uno o más órganos competentes para resolver concesiones de la convocatoria.	carácter
DIR3_ANTE_CON705	Entidad organizadora	Identificación del órgano que abre la convocatoria	carácter
CON100	ID de la convocatoria	Identificador único de cada convocatoria, asignado automáticamente por el sistema informático al registrar la convocatoria en la BDNS	carácter
REF_EXTERNA	Gestor de convocatorias	Quién es el gestor de la convocatoria	carácter
DESC_CONV	Descripción de la convocatoria en español	Cuál es la descripción de la convocatoria en español	carácter
BASEDESC_CON250	Descripción de las bases reguladoras	Texto del título del reglamento que contiene las bases reguladoras para gestionar la convocatoria	carácter
BASEURL_CON260	URL de la regulación	Enlace al sitio web que contiene el texto completo en español de las bases reguladoras	carácter
ABIERTO_CON420	Período de admisión abierto	Indica si la convocatoria mantiene abierto permanentemente el plazo de admisión de solicitudes	factor
INISOLICITUD_CON440	Fecha de inicio del plazo de solicitud	Fecha de inicio del período habilitado para admitir solicitudes	fecha
FINSOLICITUD_CON460	Fecha de finalización del plazo de solicitud	Fecha de finalización del período habilitado para admitir solicitudes	fecha
AUDAESTADO_CON490	Condición de la ayuda estatal	Indica si la ayuda de la convocatoria debe clasificarse como ADE	factor
TIPOAYUDA_CON495	Tipo de autorización ADE	Mecanismo de autorización de ayudas	factor
REGLAMENTO_CON502_50	Reglamento de exención por categoría de ayuda + Reglamento de exención por importe	Reglamento de la UE de exención de la obligación de notificación previa a la Comisión por categoría de ayuda	factor
REFERENCIA_CON515	Referencia de ayuda de la UE	Referencia asignada por la UE como identificación de la ayuda	carácter
FINALIDAD_CON540	Finalidad	Utilidad pública o interés social o promoción de un fin público que se persigue con la concesión de la subvención	factor
FINJUSTIFICACION_CON600	Fecha final de justificación	Fecha de finalización absoluta del plazo de presentación de justificaciones de cualquier concesión	fecha
NOMINATIVA_CON610	Subvención nominativa	Condición de subvención nominativa	factor
PUBLICABLE_CON620	Publicación	Estado de publicidad de las concesiones	factor

Variable	Descripción breve	Descripción adicional de la variable, si procede	Tipo
IMPACTOGENERO_CON630	Impacto de género	Califica los resultados esperados en relación con la eliminación de las desigualdades entre mujeres y hombres y el cumplimiento de los objetivos en política de igualdad	factor
FECHA_ACTUALIZACION	Momento en el que se actualizó un tercero en la base de datos	Cuál es el momento en el que se actualizó un tercero en la base de datos	fecha
PAIS_TER100	País tercero	País que genera la identificación del tercero	carácter
ID_TER110	ID del tercero	Identificación de tercero	carácter
NOMBRE_TER210_240.x	Apellido + Nombre comercial	Según la información facilitada al organismo obligado por el tercero	carácter
PAISDOM_TER250	País del tercero	Cuál es el país del tercero	factor
DOMICILIO_TER252	Dirección del tercero	Cuál es la dirección del tercero	carácter
CODPOSTAL_TER254	Código postal del tercero	Cuál es el código postal del tercero	carácter
PROVINCIA_TER258	Provincia del tercero	Cuál es la provincial del tercero	carácter
MUNICIPIO_TER256	Municipio del tercero	Cuál es el municipio del tercero	carácter
NATURALEZA_TER280	Naturaleza jurídica del tercero	Cuál es la naturaleza jurídica del tercero	factor
TIPOBEN_TER290	Tipo de tercero	Catalogación de terceros en función de su naturaleza jurídica y actividad económica	factor
REGION_TER310	Región	Región en la que se encuentra el tercero	carácter
ID_COS	ID de la concesión	Cuál es la identificación única de la concesión	carácter
TIPO_CONC_CSU204	Herramienta	Una de las figuras jurídicas o económicas en función de la cual se conceden las subvenciones y ayudas	factor
DISCRIMINADOR_CSU130	Discriminador	Discriminador de concesiones de subvenciones	carácter
OBJETIVOS_CSU205	Objetivo	Objetivo del Reglamento de exención de categorías de ayuda	factor
DAT_COS_CSU210	Fecha de concesión de subvenciones	Fecha de resolución de la concesión de la subvención	fecha
COSTE_ACT_CSU240	Costes	Importe del presupuesto financiable de la actividad para la que se solicita la subvención	numérico
IMPORTE_CONCE_CSU220	Importe concedido (subvención)	Importe total comprometido en la concesión de la subvención	numérico
AYUDA_EQUI_CSU250	Ayuda equivalente (subvención)	Ayuda equivalente de concesión de subvención	numérico
REGION_CSU260	Región	Ubicación geográfica de la solicitud material de la concesión de la subvención	carácter
Año	Año de la concesión	En qué año se concedió la concesión	fecha
CSU110	Anuncio	Identificación de la convocatoria	carácter
PAIS_CSU120	País beneficiario	Cuál es el país del beneficiario	factor
ID_CSU120	Beneficiario	Identificación del beneficiario	carácter
CSU130	Discriminador	Referencia propia de la entidad concedente, contenido libre, utilizada para discriminar cada concesión de subvención a un mismo beneficiario en la misma convocatoria	carácter
DISCRIMINADOR_PAG110	Discriminador de pagos	Referencia propia de la entidad concedente, contenido libre, utilizada para discriminar cada pago de la misma concesión	carácter
FECHAPAGO_PAG210	Fecha de pago (subvención)	Cuándo se pagó la subvención	fecha
IMPORTE_PAG220	Importe pagado (subvención)	Cuál es el importe de la concesión	numérico

Variable	Descripción breve	Descripción adicional de la variable, si procede	Tipo
RETENCION_PAG230	Retención	Condición de la retención fiscal realizada	factor
PROYECTO_PRO130	ID del proyecto	Identificación del proyecto	carácter
DESCRIPCION_PRO210	Descripción	Descripción del proyecto	carácter
IMPORTEPROY220	Importe de la subvención	Importe de la subvención asignado al proyecto	numérico
IMPORTEPROY230	Importe del préstamo	Importe del préstamo concedido al proyecto	numérico
COSTE_PRO240	Costes del proyecto	Coste financiable del proyecto	numérico
AYUDA_PRO250	Ayuda equivalente (proyecto)	Apoyo equivalente al proyecto	numérico
REGION_PROY260	Región	Ubicación geográfica del proyecto	carácter
ANIO_EJE130	Año	Año de ejecución del proyecto	fecha
EJE210	Importe de la subvención	Importe de la concesión de la subvención asignado al ejecutor en el año.	numérico
EJE220	Importe del préstamo	Importe de la concesión del préstamo asignado al ejecutor en el año.	numérico
COSTE_EJE240	Costes	Ayuda equivalente al ejecutor del proyecto en el año	numérico
AYUDA_EJE250	Ayuda equivalente (ejecutor)	Cuál es el importe de la ayuda proporcionada al ejecutor	numérico
IDENTIFICADOR_EJE120	ID del ejecutor	Identificación del ejecutor	carácter
DISCRIMINADOR_REI110_1...8	Discriminador de devoluciones	Referencia propia de la entidad concedente, contenido libre, utilizada para discriminar cada procedimiento de revolución a un mismo beneficiario procedente de la misma concesión	carácter
FECHADEV_REI210_1...8	Fecha de resolución de la devolución	Fecha de resolución del procedimiento de devolución del trámite	fecha
PRINCIPAL_REI230_1...8	Principal	Importe de la devolución	numérico
CAUSA_REI220_1...7	Causas	Una o más de las causas que sustentan el origen de la devolución	factor
DISCRIMINADOR_DEV110	Discriminador de devoluciones	Referencia propia de la entidad concedente, contenido libre, utilizada para discriminar cada devolución voluntaria de uno o más pagos de la misma concesión	carácter
FECHADEV_DEV210	Fecha de devolución	Fecha de la resolución administrativa de aceptación de la devolución.	fecha
PRINCIPAL_DEV220	Importe del principal de la devolución	Importe del principal que devuelve el beneficiario sin resolución de devolución.	numérico
INTEERESES_DEV230	Cantidad de intereses de la devolución	Cantidad de intereses de demora calculados	numérico
CON550	Actividades económicas	Uno o más de los sectores de la economía previstos en la convocatoria	factor
STRDESCRIPCION.x	Descripción de CON550	Descripción de la variable CON550	carácter
DIARIO_CON310	Diario oficial de publicación	Referencia al Boletín Oficial al que debe enviarse el extracto de la convocatoria para su publicación	carácter
DESCRIPCION_CON335	Título en español de la convocatoria	Cuál es el título en español de la convocatoria	carácter
FECHA_CON351	Fecha de la firma de la convocatoria	Cuándo se firmó la convocatoria	fecha

Variable	Descripción breve	Descripción adicional de la variable, si procede	Tipo
LOCALIDAD_CON352	Ubicación del pie de firma de la convocatoria	Dónde se firmó la convocatoria	carácter
PUBLI_CON390	Fecha de publicación en el boletín oficial	Cuándo se publicó la convocatoria en el boletín oficial	fecha
URL_CON400	Referencia en el boletín oficial del extracto en español	Cuál es el extracto en el boletín oficial español	carácter
CON560	Herramienta de ayuda	Una o más de las figuras jurídicas o económicas en función de la/s cual/es se conceden las subvenciones y ayudas	factor
STRDESCRIPCION.y	Descripción de CON560	La descripción de la variable CON503	carácter
CON503	Objetivo del Reglamento de exención de categorías de ayuda	Cuáles son los objetivos del Reglamento de exención de categorías de ayuda	factor
STRDESCRIPCION_CON503	Descripción de CON503	Descripción de la variable CON503	carácter
CON580	Tipos de beneficiarios	Uno o más de los tipos de beneficiarios previstos en la convocatoria	factor
STRDESCRIPCION_CON580	Descripción de CON580	Descripción de la variable CON580	carácter
CON570	Regiones geográficas	Una o más ubicaciones geográficas de la aplicación material de la subvención o ayuda prevista en la convocatoria	factor
STRDESCRIPCION_CON570	Descripción de CON570	Descripción de la variable CON570	carácter
CON690	Importe de financiación del fondo de la UE	Cuál es el importe de financiación de los fondos de la UE	numérico
STRDESCRIPCION_CON690	Descripción de CON690	Descripción de la variable CON690	carácter
STRVALOR	Tipo de institución financiera de la UE	Cuál es la institución financiera de la UE	factor
DATSANC_1...3x	Fecha de sanción	Cuándo se impuso la sanción	fecha
STRDISCRIMINADOR_1...3x	Discriminador de la sanción	Qué organización es la discriminadora de la sanción	carácter
MULTALEVE_SAN250_1...3x	Multa por infracciones menores	Cuál es el importe de la multa por las infracciones menores	numérico
MULTAGRAVE_SAN280_1...3x	Multa por infracciones graves	Cuál es el importe de la multa por las infracciones graves	numérico
MULTAMUYGRAVE_1...3x	Multa por infracciones muy graves	Cuál es el importe de la multa por infracciones muy graves	numérico
PUBLICABLE_SAN440_1...3x	Estado de publicidad de las sanciones	Indica si la sanción debe ser pública, según el art. 20.9 LGS	factor
LIMITE_SAN450_1...3x	Plazo de publicidad	Cuál es el plazo para publicitar	carácter
STRDESCRIPCION_SANC_1...3x	la descripción del valor STRVALOR	Descripción del valor de la variable STRVALOR	carácter
STRVALOR2_1...3x	Tipo de infracción	Comportamientos leves + Comportamientos graves + Comportamientos muy graves	factor
ACTIVIDADES_TER320	Actividades económicas de terceros	Cuáles son los tipos de actividades económicas de terceros	factor
DESC_ACTIVIDAD	Descripción de TER320	Descripción de valores de la variable TER320	carácter

Anexo C. Directorio de variables utilizadas en el análisis

Variable	Descripción breve	Descripción de la variable	Tipo
ABIERTO_CON420	Período de admisión abierto	Indica si la convocatoria mantiene abierto permanentemente el plazo de admisión de solicitudes	factor
AUDAESTADO_CON490	Condición de la ayuda estatal	Indica si la ayuda de la convocatoria debe clasificarse como ADE	factor
FINALIDAD_CON540	Finalidad	Utilidad pública o interés social o promoción de un fin público que se persigue con la concesión de la subvención	factor
NOMINATIVA_CON610	Subvención nominativa	Condición de subvención nominativa	factor
PUBLICABLE_CON620	Publicación	Estado de publicidad de las concesiones	factor
IMPACTOGENERO_CON630	Impacto de género	Califica los resultados esperados en relación con la eliminación de las desigualdades entre mujeres y hombres y el cumplimiento de los objetivos en política de igualdad	factor
PAIS_TER100	País tercero	País que genera la identificación del tercero	factor
PAISDOM_TER250	País de domicilio	País en el que está ubicado el tercero	factor
NATURALEZA_TER280	Naturaleza jurídica del tercero	La naturaleza jurídica del tercero	factor
TIPOBEN_TER290	Tipo de tercero	Catalogación de terceros en función de su naturaleza jurídica y actividad económica	factor
COSTE_ACT_CSU240	Costes	Importe del presupuesto financiable de la actividad para la que se solicita la subvención	numérico
IMPORTE_PAG220	Importe pagado (subvención)		numérico
RETENCION_PAG230	Retención	Condición de la retención fiscal realizada	factor
CON560	Herramienta de ayuda	Una o más de las figuras jurídicas o económicas en función de la/s cual/es se conceden las subvenciones y ayudas	factor
CON580	Tipos de beneficiarios	Uno o más de los tipos de beneficiarios previstos en la convocatoria	factor
SAN_dum	Sanciones	Si la concesión fue sancionada	factor
Month_CSU210	Mes de la concesión	Mes de la fecha de concesión de la subvención	factor
Nawards_TER_110	Número de concesiones	Número de concesiones recibidas por el mismo tercero	numérico
Amount_awards_TER110	Cantidad de concesiones	Cantidad total de concesiones recibidas por el mismo tercero	numérico
NATIONAL_CSU260 REGIONAL_CSU260 MUNICIAPAL_CSU260	Nivel de concesión	Si la subvención fue concedida por un organismo nacional, autonómico o municipal	factor

NATIONAL_TER310 REGIONAL_TER310 MUNICIAPAL_TER310	Nivel de ubicación del tercero	Si el tercero está ubicado a nivel nacional, autonómico, municipal	factor
LOCAL_IMPL	Implantación local	Si la ubicación del tercero es la misma que la ubicación de la entidad concedente	factor
SECTOR_CON550_AGR.. .EXTRATER	Sector de economía	Sectores de la economía previstos en la convocatoria	factor

Glosario

Algoritmo	Los algoritmos son conjuntos secuenciales exactos de comandos que se ejecutan sobre una entrada diseñada para generar una salida en un formato claramente definido. Los algoritmos se pueden representar en lenguaje sencillo, diagramas, códigos informáticos y otros lenguajes.
Beneficiario/Destinatario de la subvención/Concesionario	Cualquier persona u organización que reciba subvenciones para apoyar sus operaciones (también denominados destinatarios, beneficiarios o concesionarios)
Conflicto de intereses	Un conflicto de intereses implica un conflicto entre el deber público y los intereses privados de un funcionario público, en el que el funcionario público tiene intereses de carácter privado que podrían influir indebidamente en el desempeño de sus obligaciones y responsabilidades oficiales.
Control	Cualquier acción emprendida por la administración, la junta y otras partes para gestionar el riesgo y aumentar la probabilidad de que se logren los objetivos y metas fijadas.[1]
Corrupción	El uso indebido activo o pasivo de los poderes de los funcionarios públicos (nombrados o elegidos) para obtener beneficios financieros privados o de otro tipo
Análisis de datos	Un proceso de inspección, limpieza, transformación y modelado de datos con el objetivo de destacar información útil, sugerir conclusiones y respaldar las decisiones.
Arquitectura de datos	La arquitectura de datos está compuesta de modelos, políticas, reglas o normas que rigen qué datos se recopilan y cómo se almacenan, disponen, integran y se usan en sistemas de datos y organizaciones.
Limpieza de datos	Un conjunto de procedimientos designados para identificar y corregir, siempre que sea posible, cualquier error de datos, incoherencias y características de datos que no estén claras.
Diccionario de datos	Un catálogo de datos que detalla el contenido de una base de datos. Incluye información sobre cada campo en las tablas de atributos y sobre el formato, las definiciones y las estructuras de las tablas de atributos. Un diccionario de datos es un componente esencial de información de metadatos.
Gobernanza de datos	La gobernanza de datos es un sistema de derechos de decisión y responsabilidades para los procesos relacionados con la información, ejecutado según modelos acordados que detallan quién puede tomar qué acciones con determinada información, y cuándo, en qué circunstancias, usando determinados métodos.
Doble financiación	Una situación en la que los mismos costes para la misma actividad se financian dos veces con fondos públicos.
Control previo	Un control que pretende reducir la posibilidad de un resultado indeseado
Control posterior	Un control que pretende identificar errores, después de un evento
Fraude	El fraude es un delito económico que implica engaño, artimañas o falsas pretensiones, mediante el cual alguien gana ilegalmente. Un fraude real está motivado por el deseo de causar daño engañando a otra persona, mientras que un fraude constructivo es una ganancia obtenida de una relación de confianza.
Subvención	Las subvenciones son transferencias en efectivo, bienes o servicios para las que no se exige devolución.
Aprendizaje automático	Un subconjunto de inteligencia artificial en el que las máquinas aprovechan los enfoques estadísticos para aprender de los datos históricos y hacer predicciones en situaciones nuevas.
Apropiación indebida	Actos que implican el robo o uso indebido de los activos de una organización.
Análisis de redes	Un conjunto de técnicas integradas para identificar las relaciones entre actores y analizar las estructuras sociales o patrones que surgen de la recurrencia de estas relaciones.

***Bagging* positivo sin etiquetar/PU**	El aprendizaje positivo sin etiquetar (PU) es una técnica de aprendizaje automático semisupervisada que permite trabajar con datos muy desequilibrados. El aprendizaje PU podría utilizarse en casos en los que la mayoría de todas las observaciones disponibles pertenecen a casos sin etiquetar
Random Forests	*Random Forest* es un algoritmo de aprendizaje automático de uso común que combina la salida de varios árboles de decisión para llegar a un único resultado. Maneja problemas de clasificación y regresión.
Valores de SHAP	Los valores de SHAP (Shapley Additive exPlanations) expresan las aportaciones marginales medias de todos los pronosticadores al resultado pronosticado.
Supervisado (aprendizaje automático)	El aprendizaje supervisado es una subcategoría del aprendizaje automático y la inteligencia artificial. Se define por el uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifiquen datos o predigan resultados con precisión. A medida que los datos de entrada se introducen en el modelo, este ajusta sus pesos hasta que el modelo se ha ajustado adecuadamente.
Conjunto de datos de prueba	Una muestra seleccionada al azar del conjunto de datos que se usa para evaluar la calidad (p. ej., precisión de pronóstico) del modelo estimada en el conjunto de datos de entrenamiento.
Conjunto de datos de formación	Una muestra seleccionada al azar del conjunto de datos que se usa para estimar («entrenar») el modelo de aprendizaje automático. Los conjuntos de datos de entrenamiento y pruebas son exclusivos mutuamente; eso significa que cada observación pertenece, bien a los conjuntos de datos de entrenamiento o a los de pruebas.

[1]Instituto de Auditores Internos. (n.d.). *Gobierno, Riesgo y Control.* Recuperado de https://na.theiia.org/standards-guidance/topics/pages/governance-risk-and-control.aspx (recuperado el 2 de noviembre de 2021).

Estudios de la OCDE sobre Gobernanza Pública

La Lucha contra el Fraude en las Subvenciones Públicas en España

APRENDIZAJE AUTOMÁTICO PARA EVALUAR LOS RIESGOS Y ORIENTAR LAS ACTIVIDADES DE CONTROL

Tras la pandemia del COVID-19, los gobiernos se enfrentan a riesgos de fraude tanto antiguos como nuevos, algunos de ellos a niveles sin precedentes, relacionados con el gasto en socorro y recuperación. Los programas de subvenciones públicas son un área de alto riesgo, en la que cualquier fraude acaba desviando el dinero de los contribuyentes de las prestaciones indispensables para los particulares y las empresas. Este informe identifica cómo la Intervención General de la Administración del Estado (IGAE) podría identificar y controlar mejor los riesgos de fraude en las subvenciones. Demuestra cómo las técnicas innovadoras de aprendizaje automático pueden ayudar a la IGAE a mejorar su evaluación de los riesgos de fraude en los datos de las subvenciones. Presenta un modelo de riesgo de trabajo, desarrollado con conjuntos de datos a disposición de la IGAE y mapea conjuntos de datos que se podrían utilizar en el futuro. El informe también considera las condiciones previas para la analítica avanzada y las evaluaciones de riesgo, incluyendo las formas en que la IGAE puede mejorar su gobernanza y gestión de datos.

Cofinanciado por la Unión Europea

IMPRESA ISBN 978-92-64-52745-4
PDF ISBN 978-92-64-49803-7

9 789264 527454