Recuadro 2

EXPERIENCIAS RECIENTES DE LA CENTRAL DE BALANCES EN EL USO DE CIENCIA DE DATOS

Este recuadro contiene un resumen de dos experiencias desarrolladas por la Central de Balances (CB) para la mejora de la información existente mediante el uso de técnicas de ciencia de datos. Dichas experiencias han sido presentadas en la reunión del Irving Fisher Commitee y el *International Statistical Congress* de Ottawa, en agosto de este año (https://www.isi2023.org/conferences/ottawa-2023/).

A Imputación estadística de valores faltantes en cuestionarios de pymes

La CB recopila datos anuales de carácter contable de las sociedades individuales no financieras mediante dos fuentes distintas: una de colaboración voluntaria (CBA) y otra procedente del Registro Mercantil (CBB, según la terminología de la CB. En esta última, recibe cada año las cuentas anuales de cerca de un millón de sociedades y las pymes quedan ampliamente representadas.

Debido al elevado número de empresas que integran la base de datos CBB, cada cuestionario se somete a un conjunto de contrastes automáticos (que verifican, básicamente, que se cumplen unas reglas lógico-aritméticas) que contribuyen a depurar la información y a determinar su calidad y coherencia interna. Aproximadamente el 20 % de la información es finalmente considerada de baja calidad para su integración en estudios y agregados, debido a los exigentes filtros que se aplican. Uno de los motivos que provoca que los cuestionarios no sean perfectos ni puedan utilizarse en las estadísticas que produce la CB es la ausencia de datos para determinados desgloses de información, que impide el cuadre aritmético de la información contable.

Con el fin de obtener una mayor muestra de cuestionarios y de que esta se acerque al total poblacional, se ha realizado un proyecto de imputación de valores faltantes mediante la aplicación de técnicas de ciencia de datos. Este proyecto trata de imputar los desgloses solamente de aquellas partidas contables correspondientes al balance que tengan datos ausentes, esto es, partidas que son desgloses del activo total, el pasivo total y el patrimonio neto, como sería el caso de los activos no corrientes y su detalle, o los detalles de financiación a corto plazo con su desglose en partidas.

Para ello se han utilizado cuestionarios con balances del año 2018 perfectamente cuadrados (como conjunto de entrenamiento) y cuestionarios de 2020 (como conjunto de validación). En un primer paso, los cuestionarios de estos dos ejercicios se vacían de forma aleatoria para simular los cuestionarios con ausencia de datos.

Después se ajustan una serie de algoritmos de aprendizaje automático *(machine learning)* con la finalidad de que «aprendan» las relaciones entre las diferentes partidas contables. Se ajusta un algoritmo para cada nodo, es decir, para cada sector y tamaño de empresa, ya que se considera que el comportamiento de las partidas contables puede ser diferente dependiendo de estas dos dimensiones. Se ha trabajado con múltiples algoritmos (KNN, MICE y missRanger) y finalmente el algoritmo que mejores resultados ha ofrecido ha sido el miceRanger, un algoritmo basado en técnicas de bosque aleatorio *(random forest)*[1]. A la hora de realizar las imputaciones, se les han impuesto unos límites para no incurrir en errores graves.

Una vez obtenidas las imputaciones de estos algoritmos, se realiza un ajuste contable para cuadrar la suma de los desgloses imputados con sus sumatorios correspondientes.

Por último, las imputaciones finales se contrastan con los datos reales de las partidas contables para comprobar que los algoritmos se han imputado correctamente. Todo este proceso queda resumido en el esquema 1.

El primer objetivo del proyecto es comprobar si las estadísticas obtenidas a nivel agregado con datos reales y con datos imputados son similares. Los resultados son significativamente satisfactorios agregando por diferentes niveles (partida, sector, tamaño y nodo). En los gráficos 1 y 2 se pueden ver los resultados para la agregación de las partidas (tanto para el total activo como para el total pasivo).

En segundo lugar, para cada cuestionario se han comparado los resultados de imputación con un *benchmark* (modelo base) y se ha comprobado que los errores medios por cuestionario se reducen considerablemente.

1 Bosque aleatorio: algoritmo de aprendizaje automático supervisado que se utiliza para dar solución a problemas de clasificación y regresión. Construye árboles de decisión a partir de diferentes muestras y selecciona el voto mayoritario para decidir la clasificación y el promedio en caso de regresión.

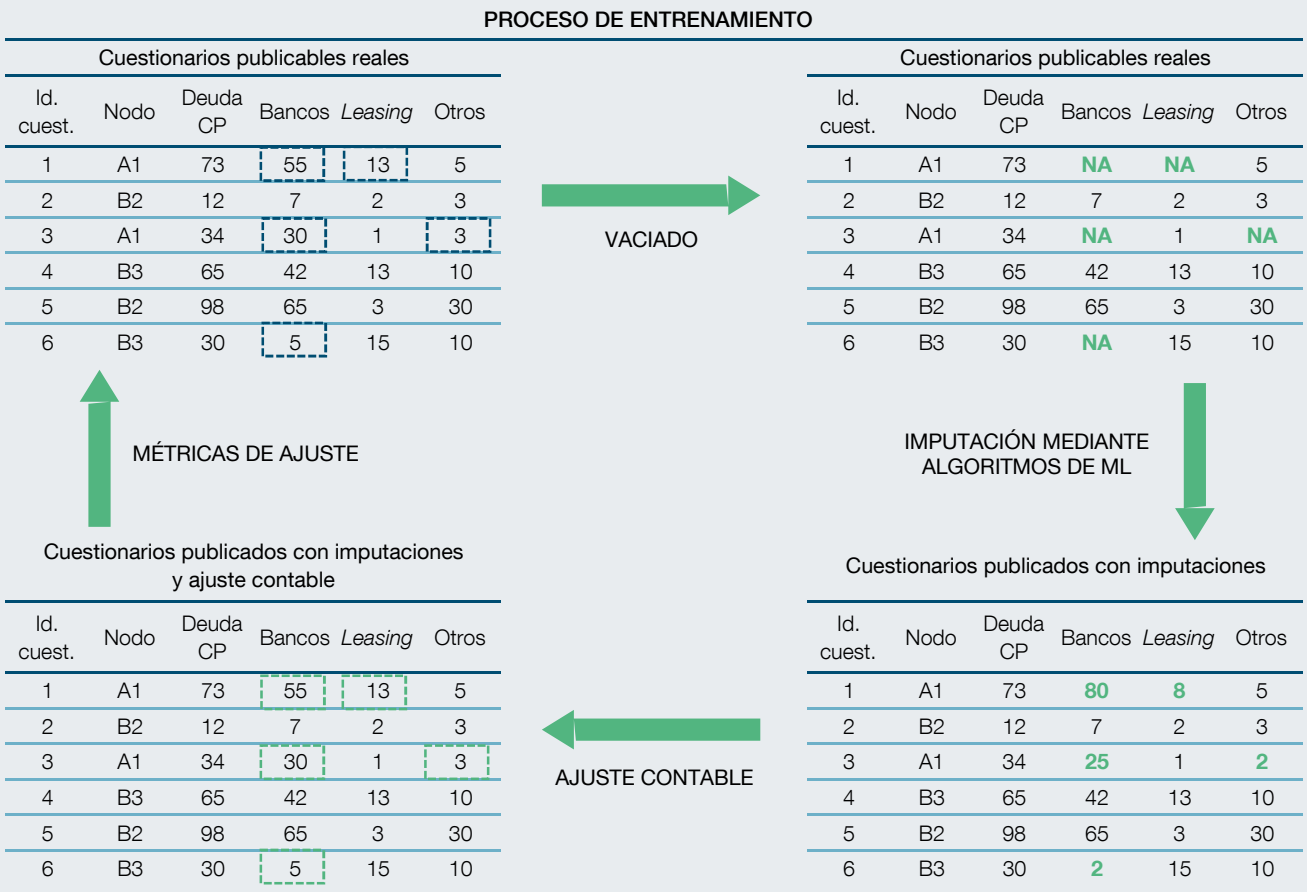
Además, se han desarrollado técnicas de explicabilidad (tanto global como local) para entender cuáles han sido las partidas contables que más han influido a la hora de realizar las diferentes imputaciones en cada cuestionario.

Una vez ajustados todos los algoritmos, se está trabajando en poner en producción este sistema, desarrollado con el lenguaje de programación de código abierto R, de tal manera que se lance de manera automática y sin intervención humana.

B Clasificación y sectorización de *holdings* y sedes centrales

La aplicación de las normas contenidas en el manual del SEC 2010 sobre sectorización de las sociedades incluye, entre otras, la distinción de las sociedades *holding*[2] dentro del sector institucional S.12. Instituciones financieras. Para facilitar la tarea de identificación de dichas sociedades, se han desarrollado de manera experimental dos trabajos consecutivos mediante la aplicación de ciencia de datos:

Esquema 1
FLUJO DE ENTRENAMIENTO DE IMPUTACIÓN



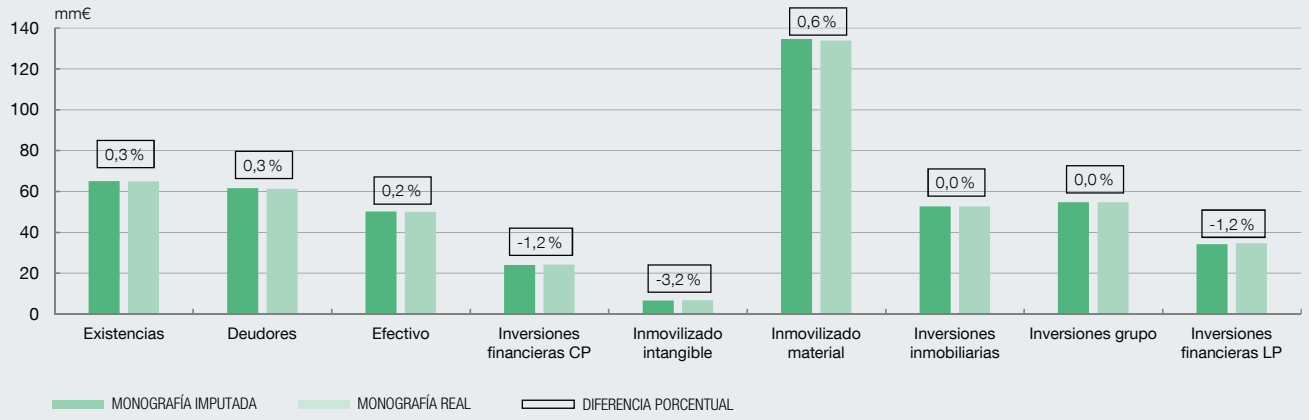
FUENTE: Banco de España.

2 Las sociedades *holding* son aquellas unidades que se dedican a la tenencia de los activos de un grupo de sociedades filiales, cuya actividad principal es la propiedad del grupo. Las sociedades *holding* no prestan ningún otro servicio a las empresas en las que mantienen una participación, es decir, no administran ni gestionan otras unidades.

a) *Asignación de la rama de producción.* En primer lugar, se ha desarrollado un procedimiento automatizado para distinguir las empresas con actividad económica (es decir, rama de producción) de *holding* y las empresas con actividad económica de sede central. El propósito es detectar empresas con posible CNAE 6420 (actividad de *holding,* esto es, meras tenedoras de cartera) o 7010 (actividad de sede central, esto es, además, gestión efectiva de la política de un grupo) mediante la verificación o no de que aquellas que declaran dichas actividades muestran indicadores (ratios económicas y financieras) que lo confirman.

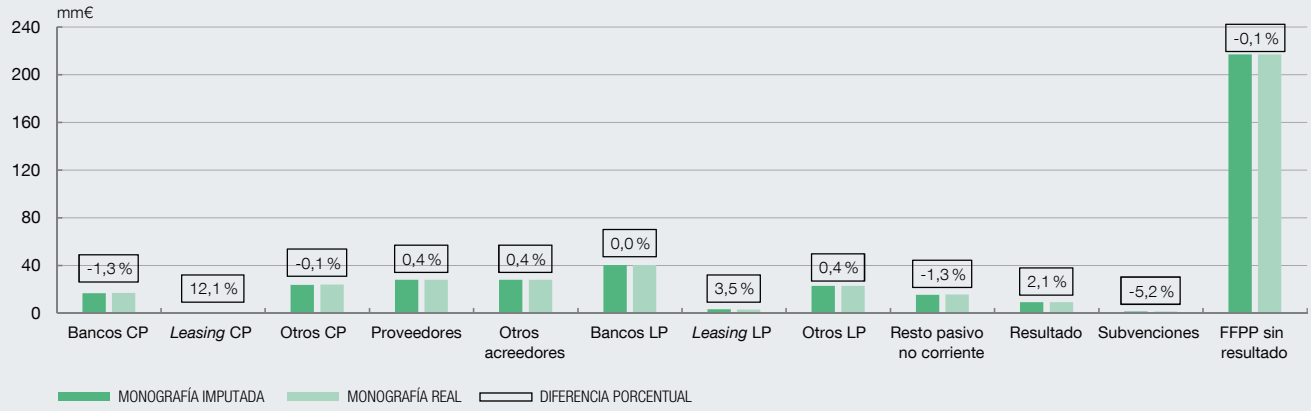
b) *Sectorización institucional diferenciando los holdings de carácter financiero.* En segundo lugar, el proyecto trata de automatizar la localización, entre las anteriores empresas catalogadas en la rama de *holdings* (CNAE 6420), de aquellas cuya sectorización institucional (es decir, la clasificación necesaria conforme a los sistemas de Cuentas Nacionales, diferente a la mera actividad económica) se encuentra en el sector financiero o en el no financiero. Esta sectorización se basa en el principio de autonomía de decisión: se consideran *holdings* de carácter financiero aquellas entidades que, por convención o por situación, tienen

Gráfico 1
DIFERENCIAS ENTRE LA IMPUTACIÓN, OBTENIDA MEDIANTE EL MÉTODO DE APRENDIZAJE AUTOMÁTICO MICERANGER, Y LOS VALORES REALES EN AGREGADO PARA LAS PARTIDAS DEL TOTAL ACTIVO DE LA MONOGRAFÍA DE LA CB PARA EL AÑO 2020



FUENTE: Banco de España.

Gráfico 2
DIFERENCIAS ENTRE LA IMPUTACIÓN, OBTENIDA MEDIANTE EL MÉTODO DE APRENDIZAJE AUTOMÁTICO MICERANGER, Y LOS VALORES REALES EN AGREGADO PARA LAS PARTIDAS DEL TOTAL PASIVO DE LA MONOGRAFÍA DE LA CB PARA EL AÑO 2020



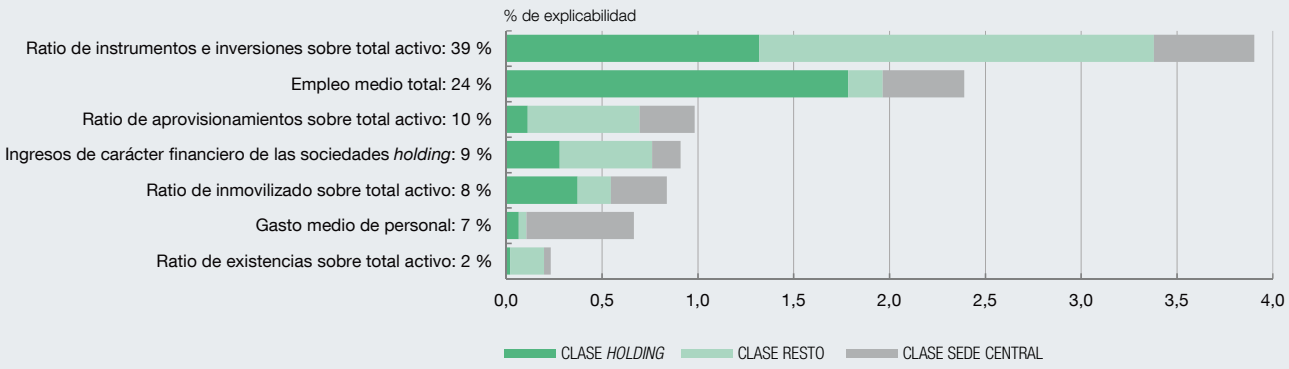
FUENTE: Banco de España.

autonomía de decisión respecto de sus socios o dueños. El sistema establece que son aquellas que cumplen una de las siguientes condiciones: a) existen varios accionistas, que harán que las decisiones de la empresa puedan ser independientes de las de su socio; b) por convención, el accionista dominante es un no residente; c) el accionista dominante está sectorizado en el sector financiero (es, por ejemplo, una sede central de una institución financiera). Para lograr esto, el modelo e información generados por la primera parte del proyecto se utilizan como punto de partida en esta segunda fase.

Para cumplir con ambas tareas, se han utilizado modelos de aprendizaje automático supervisado para clasificación. Un modelo supervisado requiere un conjunto previo de empresas etiquetadas, lo que significa que necesita empresas categorizadas con antelación y con total certeza como *holding* / sede central / otra actividad, ya sea financiera o no financiera. En las bases de datos disponibles en la CB, hay una amplia gama de empresas previamente procesadas por el personal de negocio cuya información se ha utilizado como información etiquetada. Para dar una idea del volumen de datos en juego, para el entrenamiento del primer modelo se han empleado alrededor de 1.600 entidades revisadas por el personal de negocio. Una vez generado, el modelo ha sido aplicado a casi todas las empresas de la CB (alrededor de 900.000). Para la asignación de la sectorización institucional *(holdings* financieros vs. no financieros), se han utilizado alrededor de 4.000 entidades para el entrenamiento del modelo que posteriormente se ha aplicado a la población de *holdings,* según rama de actividad (16.000 entidades por año), para determinar su sector institucional (financiero o no financiero).

Los resultados obtenidos en la aplicación de los dos modelos de IA se concretan en dos acciones independientes que está poniendo en producción la CB: por una parte, la asignación automática de más de 8.500 empresas a las ramas de actividad de *holdings* y/o sedes centrales en aquellos casos en que el resultado del modelo de rama de actividad se alinea con las reglas de negocio (por ejemplo, cuando más del 50 % del Activo está materializado en instrumentos de patrimonio de empresas del grupo), y la sugerencia de revisión manual de, aproximadamente, otras 5.300 empresas; por otra parte, la utilización del modelo de sectorización institucional para reducir el conjunto de entidades en las que hay que realizar una revisión, lo que ahorra esfuerzo a los analistas de la CB.

Gráfico 3
VARIABLES Y SU INFLUENCIA EN EL MODELO FINAL DE RAMA DE ACTIVIDAD EN LAS BASES DE DATOS ANUALES 2019 Y 2020 DE LA CB (a)



FUENTE: Banco de España.

a En el eje de ordenadas, aparecen las variables ordenadas de manera descendente por su importancia en el modelo, gracias a los valores absolutos de los valores de Shapley. Por ejemplo, la segunda variable con más influencia es el empleo medio total, con un 24 %. Esa variable, además, ayuda a distinguir las empresas que son *holding* del resto, ya que en términos generales tienen menos empleo. Esa es la razón de que la barra verde más oscura, asociada a la clase *holding,* sea tan larga en ese caso.

Con respecto a las variables utilizadas en los modelos, se muestran en las figuras los valores de Shapley[3], que dan una idea de la influencia e impacto de cada variable económica, financiera y otras que se han visto relevantes en el modelo final.

En primer lugar, en el modelo de Rama de actividad, las dos variables principales son:

— Ratio de instrumentos e inversiones sobre total activo. Es una variable que tiene en cuenta, entre otros, los instrumentos de patrimonio en empresas del grupo y el activo total. Es natural que esta variable sea influyente, por las características naturales de las empresas de tipo *holding.*

— Empleo medio. Es la segunda variable más relevante, ya que ayuda a distinguir los *holdings* del resto y de las sedes. A partir del gráfico 4, se deduce que el empleo medio es mayor en el resto de las empresas que en las sociedades *holding.*

En el modelo de Sectorización institucional para distinguir los *holdings* financieros de los no financieros, la variable más influyente es la de Total activo, pues se observa que, en general, los *holdings* no financieros son más pequeños que los financieros.

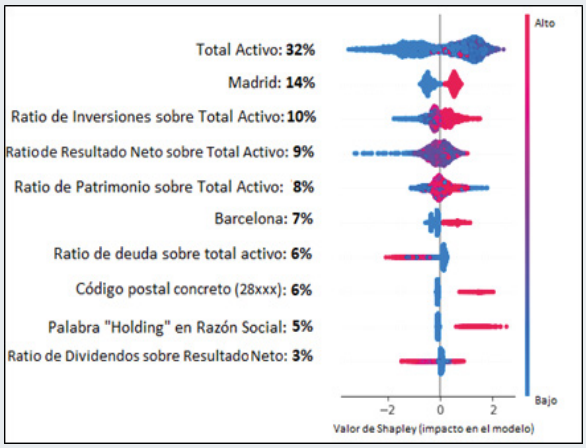
El resto de las variables se pueden clasificar del siguiente modo:

— Variables geográficas. Explican un 27 % del modelo. La mayoría de los *holdings* financieros tienen su sede en Madrid y Barcelona; incluso determinados códigos postales tienen valor explicativo por la concentración de *holdings* en ellos.

— Ratios contables, especialmente las inversiones en empresas del grupo.

— La palabra *holding* está presente en la razón social. Esta palabra aparece con más frecuencia en los *holdings* financieros que en los no financieros.

Gráfico 4
VARIABLES Y SU INFLUENCIA EN EL MODELO FINAL DE SECTORIZACIÓN INSTITUCIONAL (a)



FUENTE: Banco de España.

a Este gráfico también representa los valores de Shapley de un modelo de aprendizaje automático, en este caso, del modelo de Sectorización institucional. Debido a que solo hay dos clases (financiera y no financiera), se puede hacer una representación más detallada de los valores de Shapley; por eso, además, no están representados en valores absolutos, sino en sus valores originales. Por ejemplo, en el caso de Madrid, los valores altos de esa variable binaria (es decir, Madrid) están situados a la derecha de 0. Así pues, el hecho de que una empresa esté localizada en Madrid hace que el modelo otorgue más probabilidades de que sea un *holding* y la clasifique como tal.

3 En el contexto del aprendizaje automático, los valores de Shapley miden la importancia de cada variable aportando un promedio de su contribución marginal a las predicciones del modelo y considerando todas las combinaciones posibles de otras variables. Esta técnica permite cuantificar cuánto influye cada variable en el resultado final del modelo; ofrece una perspectiva clara de su relevancia relativa y ayuda a interpretar cómo el modelo hace sus predicciones.