Recuadro 2

USO DE INFORMACIÓN FINANCIERA Y DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA CLASIFICACIÓN DE LAS EMPRESAS POR RAMAS DE ACTIVIDAD: APLICACIÓN A LA CNAE 2025

1 Introducción

La calidad de la clasificación de las empresas constituye un elemento fundamental para el análisis económico y para la elaboración de estadísticas empresariales. Una asignación correcta y homogénea de la actividad económica permite mejorar la comparabilidad de la información, la construcción de agregados sectoriales y la interpretación de los resultados obtenidos a partir de los datos empresariales.

Esta cuestión adquiere especial relevancia en el contexto de la implantación de la Clasificación Nacional de Actividades Económicas 2025 (CNAE 2025). La nueva clasificación introduce modificaciones respecto a la CNAE 2009 y exige revisar la codificación de las empresas para garantizar que su actividad se encuadra en la categoría que mejor encaja con su realidad económica. En determinados casos, la correspondencia entre ambas clasificaciones no es unívoca, por lo que no siempre se puede realizar una asignación directa con la información disponible en la clasificación anterior.

En el marco de este proceso, las comparaciones realizadas entre la clasificación registrada en la Central de Balances y la recogida en otras fuentes de información pusieron de manifiesto discrepancias relevantes en la identificación de la actividad económica de una parte de las empresas. Estas diferencias evidenciaron la conveniencia de incorporar elementos adicionales de contraste que permitiesen reforzar la cobertura, homogeneidad y consistencia de las asignaciones realizadas.

Con este objetivo, y a partir de una línea de investigación previa de la Central de Balances para la identificación de sociedades *holding* y sedes centrales mediante técnicas de aprendizaje automático[1], se desarrolló una herramienta que amplía este enfoque al conjunto de las ramas de actividad económica. La herramienta permite inferir la actividad más probable de una empresa a partir de su información financiera y descriptiva. En la adaptación a la CNAE 2025, se ha utilizado para apoyar la clasificación de aquellos casos en los que la correspondencia entre clasificaciones no era unívoca o la información disponible no permitía una asignación directa. De este modo, ha aportado evidencia adicional que complementa la revisión experta y los procedimientos habituales de control de calidad.

2 Datos utilizados

La información que se ha utilizado procede de la Central de Balances del Banco de España, que dispone, para cada año, de información financiera y descriptiva de aproximadamente un millón de empresas. Esta base de datos proporciona una cobertura amplia del tejido empresarial y contiene información detallada sobre la estructura patrimonial, la situación financiera y los resultados de las entidades. Por ello, constituye una fuente especialmente adecuada para analizar las diferencias económicas entre ramas de actividad.

Las empresas pertenecientes a una misma actividad suelen compartir determinadas características económicas. Por ejemplo, la composición de sus ingresos y gastos, el peso de las existencias, la importancia relativa de los costes laborales o la estructura de su financiación pueden diferir de forma significativa entre ramas. Estas diferencias permiten identificar patrones que ayudan a distinguir unas actividades de otras.

La distribución de las empresas de la Central de Balances entre las distintas secciones de la CNAE presenta una heterogeneidad significativa, que refleja, en términos generales, la composición del tejido empresarial español (véase gráfico 1). Algunas actividades, como el comercio, la construcción, las actividades inmobiliarias, las actividades profesionales o la industria manufacturera, concentran un elevado número de empresas, mientras que otras ramas cuentan con una representación considerablemente menor. Esta diversidad sectorial constituye un aspecto relevante tanto para el proceso de entrenamiento como para la interpretación de los resultados.

Para desarrollar el análisis fue necesario construir un conjunto de empresas cuya información financiera (balances y cuentas de resultados) y cuya clasificación según la actividad económica pudieran considerarse especialmente fiables y carentes de errores de información. Para ello se combinaron empresas revisadas por especialistas de la Central de Balances con otras respecto a las cuales distintas fuentes de información mostraban concordancia. Este grupo de empresas sirvió de referencia para identificar las características económicas asociadas a cada actividad.

1 Alejandro Morales Fernández. (2024). "Business Sector Classification and Beyond Using Machine Learning". Notas Estadísticas, 18, Banco de España.

Recuadro 2

USO DE INFORMACIÓN FINANCIERA Y DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA CLASIFICACIÓN DE LAS EMPRESAS POR RAMAS DE ACTIVIDAD: APLICACIÓN A LA CNAE 2025 (cont.)

La combinación de ambos criterios permitió obtener una muestra amplia y sólida para el desarrollo y evaluación de la herramienta. Su distribución por ramas de actividad reproduce, en términos generales, la estructura observada en el conjunto de empresas de la Central de Balances y proporciona una base adecuada para analizar la capacidad de diferenciación de la información financiera.

3 Funcionamiento del modelo

La idea que sustenta el proyecto es sencilla: si las empresas pertenecientes a una misma actividad económica presentan características financieras similares, esa información puede utilizarse para ayudar a identificar la actividad que desarrollan otras empresas.

Para ello se emplearon técnicas de aprendizaje automático capaces de analizar simultáneamente un gran número de variables y de identificar patrones recurrentes en la información disponible. El objetivo no era sustituir el análisis experto, sino aprovechar el volumen de información existente para detectar regularidades que contribuyeran a mejorar la clasificación de las empresas.

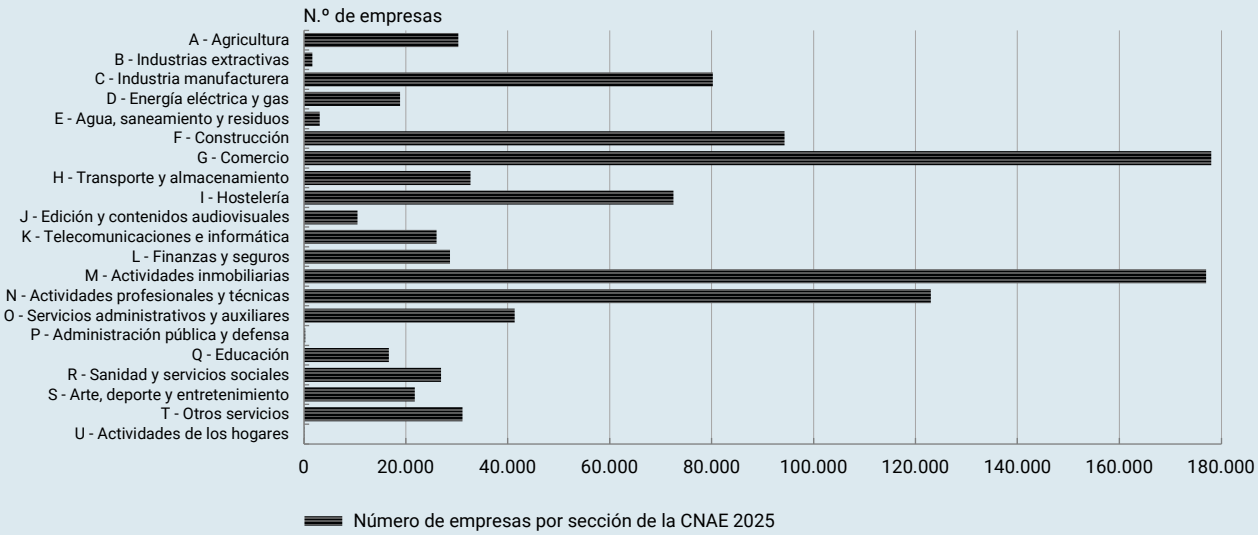
La herramienta se construyó utilizando la información contenida en los estados financieros de las empresas. En particular, se emplearon variables procedentes del balance y de la cuenta de pérdidas y ganancias, junto con un amplio conjunto de indicadores elaborados a partir de dicha información. Estos indicadores recogen aspectos relacionados con la rentabilidad, la estructura financiera, los márgenes empresariales, la composición de los activos y pasivos, el empleo y otras características relevantes de la actividad de las empresas.

La información financiera se complementó con determinadas variables descriptivas, como la localización geográfica, la razón social o la actividad declarada. En conjunto, todos estos elementos permiten obtener una visión más completa de cada empresa y facilitan la diferenciación entre actividades con características similares.

A partir de esta información, la herramienta identifica las características más representativas de cada actividad económica y estima la clasificación más probable para aquellas empresas cuya asignación a una actividad plantea dudas o para aquellos casos en los que la adaptación a la CNAE 2025 no puede realizarse de forma directa. Cada propuesta de clasificación se acompaña de una medida de confianza que permite valorar la consistencia del resultado obtenido y priorizar los casos que requieren una revisión más detallada.

Es importante señalar que la clasificación propuesta no se basa en una interpretación directa de las descripciones

Gráfico 1
Distribución de empresas por sección (CNAE 2025)



FUENTE: Elaboración propia.

Recuadro 2

USO DE INFORMACIÓN FINANCIERA Y DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA CLASIFICACIÓN DE LAS EMPRESAS POR RAMAS DE ACTIVIDAD: APLICACIÓN A LA CNAE 2025 (cont.)

oficiales de la actividad productiva aportadas por las empresas, sino en la información económico-financiera observada en empresas cuyas variables económicas y actividad principal según la CNAE se consideran suficientemente fiables. De este modo, la propuesta se apoya en los patrones económicos, financieros y descriptivos identificados en las empresas de referencia.

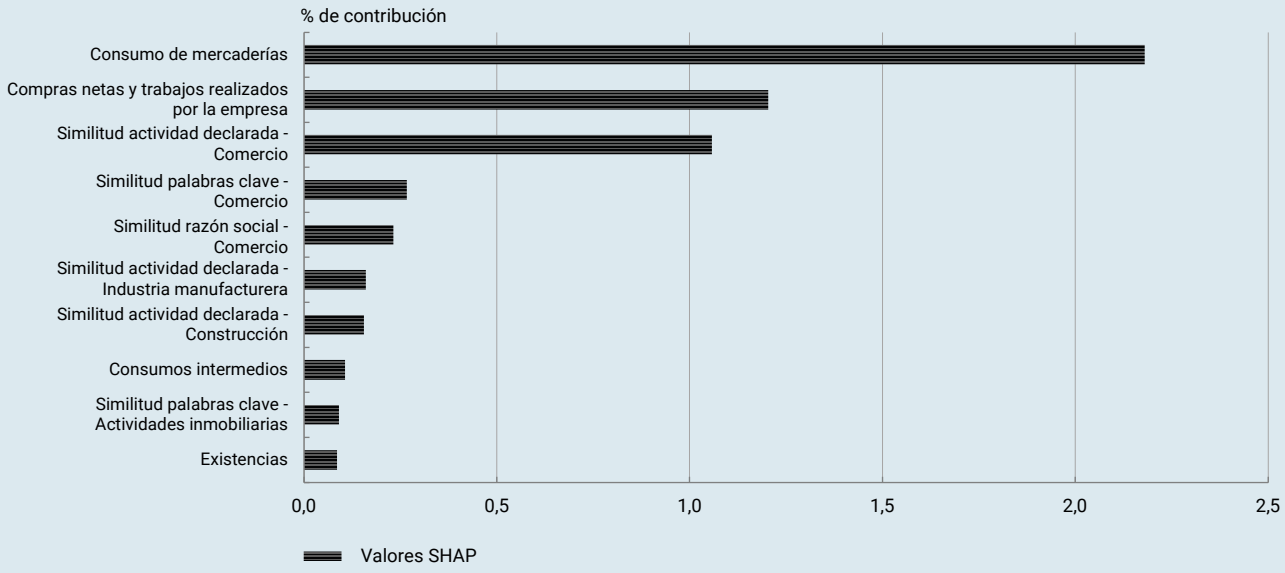
La herramienta se evaluó con empresas que no habían intervenido en su construcción, para comprobar su capacidad para clasificar nuevos casos. Además, dado que el número de empresas difiere de forma significativa entre ramas, tal como se ha visto en el gráfico 1, los resultados se analizaron tanto en conjunto como de manera separada para cada actividad.

Con el fin de comprender mejor las clasificaciones propuestas, también se analizaron las variables que tenían una mayor influencia en los resultados. Este ejercicio permitió verificar que los patrones identificados eran coherentes con las características económicas propias de las distintas ramas de actividad. Por ejemplo, en el comercio al por mayor y al por menor destacan variables vinculadas a la compra y venta de mercancías, mientras que en otras actividades resultan más relevantes factores relacionados con el empleo, la inversión o la estructura de costes. El gráfico 2 informa de las ratios o variables que han resultado más significativas en la evaluación de una rama concreta.

4 Resultados

La aplicación de la herramienta durante el proceso de adaptación a la CNAE 2025 puso de manifiesto que la información financiera y descriptiva disponible en la Central de Balances contiene señales informativas para apoyar la clasificación de las empresas por actividad económica. En particular, permitió aportar evidencia adicional para aquellos casos en los que la correspondencia entre clasificaciones no era unívoca o se arrojaban dudas sobre la asignación inicial. Los resultados son especialmente favorables en aquellas actividades que presentan características económicas diferenciadas y cuentan con una representación suficiente en el conjunto de empresas utilizado como referencia (véase cuadro 1). Entre ellas se encuentran la agricultura, la industria manufacturera, el suministro de energía, la construcción, el comercio, el transporte y almacenamiento, la hostelería,

Gráfico 2
Variables con mayor contribución a la clasificación de la sección G, «Comercio al por mayor y al por menor»



FUENTE: Elaboración propia.
NOTA: Las variables de similitud miden el grado de afinidad entre la información textual de la empresa analizada y la observada en las empresas de cada sección de la CNAE utilizadas durante el entrenamiento del modelo. La similitud de la actividad declarada utiliza la descripción de la actividad económica facilitada por la empresa; la similitud de la razón social utiliza su denominación, y la similitud de las palabras clave se centra en los términos más característicos contenidos en dicha denominación.

Recuadro 2

USO DE INFORMACIÓN FINANCIERA Y DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA CLASIFICACIÓN DE LAS EMPRESAS POR RAMAS DE ACTIVIDAD: APLICACIÓN A LA CNAE 2025 (cont.)

las actividades profesionales y las actividades sanitarias. En varias de estas ramas, la proporción de empresas correctamente clasificadas supera el 90 %[2] dentro de la muestra evaluada.

Las principales dificultades aparecen en aquellas actividades que presentan características similares a las de otras ramas o que cuentan con una representación más reducida dentro de la muestra. En estos casos, la capacidad de diferenciación del modelo es más limitada y las discrepancias suelen concentrarse entre actividades próximas desde un punto de vista económico, lo que resulta coherente con la propia naturaleza de dichas ramas. Por ello, aunque los distintos ejercicios de evaluación arrojan resultados satisfactorios para una parte importante de las actividades, las propuestas deben interpretarse teniendo en cuenta las particularidades de cada actividad y el grado de confianza asociado a cada clasificación.

En conjunto, la evidencia obtenida confirma que las técnicas de aprendizaje automático constituyen una herramienta útil para reforzar la calidad de la clasificación según la CNAE 2025, especialmente por ramas de actividad. El modelo aporta una señal adicional para la toma de decisiones, pero no sustituye la aplicación de criterios estadísticos, contables y económicos ni la revisión experta de los casos que presentan mayor complejidad.

Más allá de los resultados obtenidos, el principal interés del proyecto radica en su aplicación práctica. Durante el proceso de adaptación a la CNAE 2025, la herramienta se

Cuadro 1
Matriz de confusión normalizada sobre el F1-*score* para la CNAE 2025

Real/predicha	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A - Agricultura	0,95	0	0,01	0	0	0	0	0	0	0	0	0	0,01	0,01	0,01	0	0	0	0	0	0
B - Industrias extractivas	0,02	0,61	0,17	0,01	0,01	0,05	0,01	0	0	0	0	0	0,02	0,03	0,03	0	0	0,01	0,01	0	0
C - Industria manufacturera	0	0	0,92	0	0	0,03	0,01	0	0	0	0	0	0	0,01	0,01	0	0	0	0,01	0	0
D - Energía eléctrica y gas	0	0	0	0,95	0	0,01	0,01	0	0	0	0	0	0,01	0,01	0	0	0	0	0	0	0
E - Agua, saneamiento y residuos	0,01	0	0,06	0	0,77	0,03	0,01	0,02	0,01	0	0	0	0	0,02	0,04	0	0	0,01	0	0,01	0
F - Construcción	0	0	0,03	0	0	0,92	0	0	0	0	0	0	0,02	0,01	0,01	0	0	0	0	0	0
G - Comercio	0	0	0	0	0	0	0,97	0	0	0	0	0	0	0,01	0,01	0	0	0	0	0	0
H - Transporte y almacenamiento	0	0	0,01	0	0	0	0	0,94	0	0	0	0	0,01	0,01	0,02	0	0	0	0	0	0
I - Hostelería	0	0	0	0	0	0	0	0	0,94	0	0	0	0,03	0	0	0	0	0	0	0	0
J - Edición y contenidos audiovisuales	0	0	0,01	0	0	0	0	0	0	0,86	0,03	0	0	0,04	0,01	0	0	0,02	0	0	0
K - Telecomunicaciones e informática	0	0	0,01	0	0	0,01	0	0	0	0,01	0,87	0	0	0,06	0,02	0	0	0	0,01	0	0
L - Finanzas y seguros	0	0	0	0	0	0	0,01	0	0	0	0	0,81	0,02	0,13	0,01	0	0	0	0	0	0
M - Actividades inmobiliarias	0	0	0	0	0	0	0,01	0,01	0,01	0	0	0	0,95	0,01	0,01	0	0	0	0	0	0
N - Actividades profesionales y técnicas	0	0	0,01	0	0	0,01	0	0	0	0	0,01	0,02	0,02	0,89	0,02	0	0	0	0	0	0
O - Servicios administrativos y auxiliares	0,01	0	0,01	0	0	0,02	0,01	0,01	0,01	0	0,01	0,01	0,02	0,07	0,80	0	0	0,01	0,01	0	0
P - Administración pública y defensa	0	0	0,04	0	0,01	0,04	0,02	0,01	0,01	0,01	0,02	0,01	0,02	0,28	0,15	0,03	0,02	0,15	0,19	0,01	0
Q - Educación	0	0	0	0	0	0	0	0	0	0	0	0	0	0,02	0,01	0	0,92	0,01	0,03	0	0
R - Sanidad y servicios sociales	0	0	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0,94	0	0	0
S - Arte, deporte y entretenimiento	0	0	0	0	0	0,01	0	0	0,01	0,01	0	0	0,01	0,02	0,02	0	0,01	0	0,88	0,01	0
T - Otros servicios	0	0	0,02	0	0	0,01	0,03	0	0	0	0,01	0	0,01	0,02	0,02	0	0	0,01	0,01	0,86	0
U - Actividades de los hogares	0	0	0,03	0	0	0	0,03	0	0	0	0,03	0	0,02	0,11	0,10	0	0	0	0,02	0,02	0,64

FUENTE: Elaboración propia.

2 La evaluación del modelo se basa principalmente en el F1-*score* ponderado *(Weighted* F1), ya que esta métrica tiene en cuenta tanto la precisión como la exhaustividad de las predicciones y, además, pondera la contribución de cada sección de la CNAE en función de su tamaño muestral. Este enfoque resulta especialmente apropiado en contextos con distribuciones desbalanceadas, donde la exactitud global puede ofrecer una visión excesivamente optimista del rendimiento del modelo.

Recuadro 2

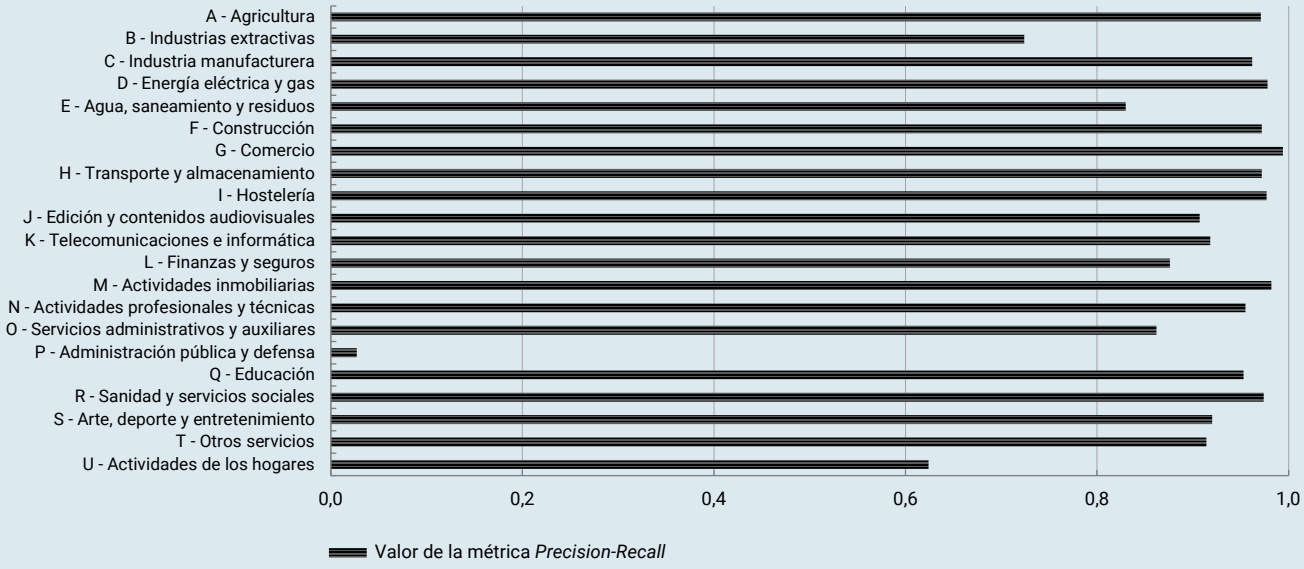
USO DE INFORMACIÓN FINANCIERA Y DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA CLASIFICACIÓN DE LAS EMPRESAS POR RAMAS DE ACTIVIDAD: APLICACIÓN A LA CNAE 2025 (cont.)

utilizó para apoyar la clasificación de un elevado número de empresas en aquellas situaciones en las que la correspondencia entre clasificaciones no era directa, no se disponía de un código CNAE suficientemente determinado o existían dudas sobre la actividad asignada. La clasificación propuesta y su medida de confianza (en el gráfico 3 esta se muestra a través de las curvas de *Precision-Recall)* proporcionaron una fuente adicional de evidencia que facilitó la resolución de casos ambiguos, la identificación de posibles inconsistencias y la priorización de las revisiones expertas.

De este modo, la información financiera de la Central de Balances se utilizó como una fuente adicional de conocimiento para apoyar la revisión de las clasificaciones existentes y mejorar la cobertura, homogeneidad y consistencia de las asignaciones realizadas. La experiencia obtenida durante la adaptación a la CNAE 2025 muestra que este enfoque aporta evidencia para reforzar los procesos de control de calidad y apoyar la identificación de las actividades económicas de las empresas.

La implantación de estos resultados, su cuantificación y las modificaciones incorporadas a los criterios de clasificación de empresas en la Central de Balances se describen con mayor detalle en el recuadro 1, «Actualización de los criterios de clasificación de empresas en la Central de Balances: revisión en profundidad y adaptación a la CNAE 2025».

Gráfico 3
Curvas de *Precision-Recall* por sección de la CNAE 2025



FUENTE: Elaboración propia.