COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

La salud financiera de los hogares españoles: evidencia a partir de la Encuesta de Presupuestos Familiares

Nombre: María Sarrado Vázquez

Director: Manuel Alejandro Betancourt Odio

MADRID | Abril 2026

Resumen

Este Trabajo de Fin de Grado analiza la salud financiera de los hogares españoles desde una perspectiva estructural y multidimensional. El estudio parte de una limitación central en la literatura aplicada y en el análisis institucional: los indicadores unidimensionales de renta no capturan con precisión la vulnerabilidad financiera efectiva de los hogares. Un hogar puede estar por encima del umbral de pobreza monetaria y, sin embargo, mantener una estructura de gasto rígida, baja capacidad de ahorro y una exposición elevada a perturbaciones económicas.

Para abordar este problema, se construye y valida una dimensión de capacidad de ahorro basada en microdatos de la Encuesta de Presupuestos Familiares (Instituto Nacional de Estadística). Metodológicamente, el trabajo aplica algoritmos de aprendizaje automático no supervisado: en primer lugar, se construye el ratio ahorro/ingresos y se depura la muestra; posteriormente, se identifican perfiles homogéneos mediante el algoritmo de clustering K-means, lo que permite segmentar a los hogares en función de su comportamiento financiero.

Los resultados muestran una heterogeneidad significativa en el comportamiento financiero de los hogares. La segmentación identifica tres perfiles diferenciados: déficit estructural, ahorrador moderado y alta capacidad financiera relativa El análisis empírico pone de manifiesto que más del 50% de los hogares presenta valores positivos del ratio ahorro/ingresos, concentrados principalmente en el intervalo entre 0 y 0,6, lo que indica la existencia de superávit presupuestario en una proporción relevante de la muestra. No obstante, una proporción menor, aunque significativa (inferior al 50%), presenta ratios negativos, reflejando situaciones de déficit estructural persistente. Asimismo, el modelo predictivo presenta un alto grado de acierto en la clasificación, como se observa en la matriz de confusión, especialmente en la identificación de los perfiles extremos. Los resultados evidencian diferencias estructurales claras entre los perfiles identificados, con centroides diferenciados en términos de ingresos, gasto y capacidad de generación de excedente, lo que confirma la existencia de patrones heterogéneos de comportamiento presupuestario.

Palabras clave: salud financiera, vulnerabilidad, ahorro, Encuesta de Presupuestos Familiares, segmentación, aprendizaje automático.

Abstract

This dissertation analyzes the financial health of Spanish households from a structural and multidimensional perspective. It addresses a key limitation in empirical research and institutional analysis: income-based indicators do not accurately capture households' effective financial vulnerability. A household may be above the poverty threshold and still exhibit a rigid expenditure structure, low savings capacity, and high exposure to economic shocks.

To address this issue, the study develops and validates a savings-capacity dimension using microdata from the Spanish Household Budget Survey (National Statistics Institute). Methodologically, the research applies unsupervised machine learning algorithms: first, a savings-to-income ratio is constructed and the sample is cleaned; subsequently, homogeneous profiles are identified using the K-means clustering algorithm, allowing households to be segmented according to their financial behavior.

The results reveal significant heterogeneity in household financial behavior. The segmentation identifies three distinct profiles: structural deficit, moderate saver, and relatively high financial capacity. The empirical analysis shows that more than 50% of households present positive values of the savings-to-income ratio, mainly concentrated between 0 and 0.6, indicating the presence of a budget surplus in a substantial proportion of the sample. However, a smaller but still significant proportion (below 50%) exhibits negative ratios, reflecting situations of persistent structural deficit. Furthermore, the results show clear structural differences between the identified profiles, with distinct centroids in terms of income, expenditure, and surplus generation capacity, confirming the existence of heterogeneous patterns of budgetary behavior.

Keywords: financial health, household vulnerability, savings, Household Budget Survey, segmentation, machine learning.

Índice de la memoria

Capítulo 1. Introducción 4
1.1 Contexto 4
1.2 Objetivos 7
1.2.1 Objetivo general 7
1.2.2 Objetivos específicos 8
1.3 Organización de la memoria 9

Capítulo 2. Marco teórico 10
2.1 Medición de la pobreza y sus limitaciones 10
2.2 Salud financiera y resilencia económica 11
2.3 Comportamiento de ahorro y nivel de ingresos 12
2.4 Aplicación de técnicas de aprendizaje automático en análisis socioeconómico 14

Capítulo 3. Metodología 16
3.1 Marco teórico de las técnicas metodológicas empleadas 16
3.1.1 Fundamentos teóricos del aprendizaje automático en el análisis socioeconómico 17
3.1.2 Aprendizaje no supervisado: el clustering como técnica de segmentación 18
3.1.3 El algoritmo K-means: fundamentos, propiedades y limitaciones 19
3.1.4 Determinación del número óptimo de clusters: el método del codo 21
3.1.5 Aprendizaje supervisado: fundamentos teóricos de los árboles de decisión 22
3.2 Fuente de datos y muestra 23
3.2.1 Encuesta de Presupuestos Familiares (EPF): caracterización y fundamento teórico 23
3.2.2 Selección muestral y proceso de depuración 24
3.3 Construcción de variables 26
3.3.1 Distribución del ratio y nivel de ahorro 27
3.4 Segmentación de hogares mediante clustering 29
3.5 Evaluación del modelo 30

Capítulo 4. ANÁLISIS DE RESULTADOS 31
4.1 Segmentación en perfiles de ahorro 31
4.2 Caracterización de los perfiles 34
4.3 Número óptimo de clusters y modelo predictivo 39

Capítulo 5. Discusión 41

5.1 Relación con los objetivos y la literatura 41

5.2 Implicaciones y limitaciones 43

Capítulo 6. Conclusiones 45

Capítulo 7. Bibliografía 48

ANEXO I 50

Fragmento de código relevante (Python) 50

Documentación y referencias estadísticas 52

Índice de figuras

Figura 1 - Distribución del ratio ahorro/ingresos tras el proceso de depuración.....**¡Error! Marcador no definido.**
Figura 2 - Distribución por nivel de ahorro (barras) ..... **¡Error! Marcador no definido.**
Figura 3 - Boxplot del ratio ahorro/ingresos por nivel de ahorro**¡Error! Marcador no definido.**
Figura 4 - Segmentación de hogares por perfil de ahorro (K-means) 32
Figura 5 - Ingresos anuales frente a gastos anual: equilibrio presupuestario 33
Figura 6 - Segmentación y centroides de los perfiles de ahorro..... 35
Figura 7 - Distribución de hogares por perfil (barras)..... 36
Figura 8 - Medias de ingresos, gasto y ratio ahorro/ingresos por perfil..... 36
Figura 9 - Boxplot del ratio ahorro/ingresos por perfil 37
Figura 10 - Proporción de hogares por perfil (diagrama de sectores) 37
Figura 11 - Ingresos anuales frente a ahorro, coloreado por perfil..... 38
Figura 12 - Método del codo para elegir el número de clusters (K-means) 39
Figura 13 - Matriz de confusión del modelo de árbol de decisión (prefil de ahorro)..... 40
Figura 14 - Resultados arbol de decisión..... 52

Índice de ecuaciones

Ecuación 1 - Criterio de optimización de K-means..... **¡Error! Marcador no definido.**
Ecuación 2 - Centroide más cercano ..... **¡Error! Marcador no definido.**
Ecuación 3 - Media de las observaciones..... **¡Error! Marcador no definido.**
Ecuación 4 - Índice de Gini ..... **¡Error! Marcador no definido.**
Ecuación 5 - Entropía ..... **¡Error! Marcador no definido.**
Ecuación 6 - Ratio ahorro e ingresos..... **¡Error! Marcador no definido.**

Capítulo 1. INTRODUCCIÓN

1.1 CONTEXTO

En los últimos años de esta última década, la estabilidad financiera de los hogares se ha consolidado como una cuestión central tanto para la política económica como para el sector asegurador y financiero. Las sucesivas crisis —financiera (2008), sanitaria (COVID-19), energética (2022) y el reciente episodio inflacionario— han puesto de manifiesto la vulnerabilidad estructural de una parte significativa de la población europea. Estas perturbaciones no solo han afectado al nivel de ingresos de los hogares, sino también a su capacidad de adaptación ante entornos económicos cambiantes, evidenciando que la estabilidad financiera depende de factores más complejos que la renta disponible.

En este contexto, la capacidad de los hogares para generar excedente financiero, afrontar gastos No imprevistos y mantener un equilibrio presupuestario estable se convierte en un elemento clave. se trata únicamente de cuánto ingresa un hogar, sino de si su estructura económica le permite sostener su nivel de vida sin recurrir de forma sistemática al endeudamiento o a la descapitalización. Esta perspectiva introduce una visión dinámica de la estabilidad financiera, en la que el comportamiento presupuestario adquiere un papel central.

Sin embargo, los indicadores tradicionales de pobreza, basados principalmente en niveles absolutos o relativos de renta, han mostrado limitaciones relevantes para capturar esta realidad. Un hogar puede situarse por encima del umbral de pobreza y, sin embargo, presentar una estructura presupuestaria frágil, caracterizada por elevados gastos fijos, baja capacidad de ahorro o escasa resiliencia ante shocks económicos. De forma análoga, existen hogares con niveles de renta moderados que mantienen comportamientos financieramente sostenibles gracias a una gestión eficiente de sus recursos. Por tanto, la mera clasificación por ingresos resulta insuficiente para evaluar la verdadera situación financiera de los hogares.

La evidencia empírica respalda estas limitaciones. Según Eurostat (2023),[1] el 21,6% de la población de la Unión Europea se encuentra en riesgo de pobreza o exclusión social, lo que equivale a aproximadamente 95 millones de personas. En el caso de España, esta cifra asciende al 26,5%, situándose de forma sistemática por encima de la media europea. Asimismo, el Banco de España (2023)[2] señala que aproximadamente un 35% de los hogares españoles no puede afrontar un gasto imprevisto de 1.000 euros sin recurrir a endeudamiento o ayuda externa. Este dato resulta especialmente relevante, ya que pone de manifiesto que la vulnerabilidad financiera no depende exclusivamente del nivel de renta, sino de la capacidad estructural del hogar para generar ahorro y absorber perturbaciones.

En este sentido, la literatura reciente ha evolucionado hacia el concepto de "salud financiera", entendido como la capacidad de los hogares para cumplir con sus obligaciones financieras, absorber shocks económicos y mantener estabilidad intertemporal. Este enfoque introduce una dimensión más amplia que la pobreza monetaria tradicional, incorporando elementos como la gestión del flujo de caja, el nivel de endeudamiento, la planificación financiera y, especialmente, la capacidad de ahorro. Sin embargo, pese a este avance conceptual, persiste un vacío en la operacionalización empírica de estos conceptos, especialmente en lo que respecta a la construcción de indicadores sintéticos que permitan medir la salud financiera de forma cuantitativa y aplicable.

A partir de esta limitación surge la necesidad de desarrollar enfoques multidimensionales que integren distintas dimensiones estructurales del comportamiento económico de los hogares. En particular, resulta relevante incorporar variables que capturen el equilibrio presupuestario, más allá de los niveles absolutos de ingresos. Este vacío es especialmente relevante tanto desde una perspectiva académica, en el ámbito de la medición de la vulnerabilidad financiera, como desde una perspectiva empresarial, donde la segmentación de clientes y la evaluación del riesgo requieren herramientas más precisas y basadas en datos.

En este contexto, la presente investigación se plantea la siguiente pregunta:

¿Cómo puede medirse de forma más precisa la salud financiera de los hogares españoles mediante un enfoque multidimensional, y en qué medida la capacidad de ahorro permite identificar y predecir perfiles de vulnerabilidad financiera?

Para dar respuesta a esta cuestión, el trabajo, desarrollado en colaboración con MAPFRE, propone el diseño y validación de un modelo cuantitativo de salud financiera basado en microdatos de la Encuesta de Presupuestos Familiares (INE). Este modelo integra distintas dimensiones económicas estructurales y utiliza técnicas de aprendizaje automático, K-means clustering y árboles de decisión, para segmentar hogares en perfiles homogéneos y evaluar su comportamiento financiero.

Dentro de este marco, la capacidad de ahorro se configura como una dimensión central del análisis. A diferencia de los indicadores basados exclusivamente en la renta, el ahorro incorpora simultáneamente ingresos y estructura de gasto, ofreciendo una visión más completa del comportamiento financiero de los hogares. Dos hogares con igual nivel de renta pueden presentar situaciones radicalmente distintas en términos de estabilidad económica si su estructura de gasto difiere sustancialmente. En este sentido, el ahorro actúa como una variable síntesis del comportamiento presupuestario, reflejando la capacidad del hogar para generar excedente y sostener su estabilidad en el tiempo.

Desde una perspectiva económica, la generación de excedente presupuestario puede interpretarse como un indicador de estabilidad intertemporal, mientras que la existencia de déficits persistentes refleja vulnerabilidad estructural. Esta interpretación permite superar el enfoque estático de la pobreza monetaria y avanzar hacia una comprensión más dinámica de la salud financiera, en la que el equilibrio entre ingresos y gastos adquiere un papel fundamental.

A partir de esta perspectiva, el objetivo general del trabajo es diseñar y validar un modelo multidimensional de salud financiera de los hogares españoles que permita segmentar perfiles financieros y predecir su comportamiento mediante técnicas de aprendizaje automático. De este objetivo general se derivan varios objetivos específicos: la construcción de indicadores estructurales representativos de distintas dimensiones

económicas, la aplicación de técnicas de segmentación no supervisada para identificar perfiles homogéneos y el desarrollo de modelos predictivos que permitan anticipar dichos perfiles a partir de variables estructurales, demográficas y territoriales.

La principal contribución del trabajo radica en integrar rigor académico y aplicabilidad empresarial, desarrollando una herramienta que no solo mejora la medición de la vulnerabilidad financiera, sino que también resulta útil para la toma de decisiones en el ámbito asegurador. En particular, el modelo permite mejorar la segmentación de clientes, la identificación de perfiles de riesgo y el diseño de estrategias adaptadas a distintos niveles de estabilidad financiera.

El resto del documento se estructura de la siguiente manera. El capítulo 2 presenta el marco teórico, abordando la medición de la pobreza, el concepto de salud financiera y la aplicación de técnicas de aprendizaje automático en el análisis socioeconómico. El capítulo 3 describe la metodología, incluyendo la fuente de datos, la construcción de variables y el diseño del modelo. El capítulo 4 expone los resultados del análisis empírico. El capítulo 5 discute los hallazgos en relación con la literatura y señala las principales limitaciones. Finalmente, el capítulo 6 recoge las conclusiones y las implicaciones del estudio.

1.2 OBJETIVOS

1.2.1 OBJETIVO GENERAL

El objetivo principal de este trabajo es diseñar y validar un modelo cuantitativo de salud financiera de los hogares españoles, basado en un enfoque multidimensional que integre distintas dimensiones económicas estructurales, con el fin de construir un indicador sintético capaz de segmentar perfiles financieros y predecir su comportamiento mediante técnicas de aprendizaje automático.

Este objetivo responde a una doble motivación: por un lado, superar las limitaciones de los indicadores tradicionales de pobreza basados exclusivamente en la renta; y por otro,

desarrollar una herramienta aplicable al ámbito asegurador que permita mejorar la segmentación de clientes y la gestión del riesgo.

1.2.2 OBJETIVOS ESPECÍFICOS

Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos:

- **Construcción de un marco multidimensional de salud financiera:** definir y operacionalizar un conjunto de dimensiones económicas estructurales que permitan capturar la estabilidad financiera de los hogares más allá del nivel de ingresos, incorporando variables relacionadas con la estructura de gasto, el esfuerzo económico y la capacidad de generación de excedente.
- **Desarrollo de un indicador de capacidad de ahorro como dimensión clave:** construir un indicador relativo de capacidad de ahorro que permita medir el equilibrio presupuestario de los hogares, garantizando su comparabilidad entre distintos niveles de renta y aplicando criterios de depuración estadística que aseguren su robustez.
- **Identificación de perfiles de comportamiento financiero:** aplicar técnicas de segmentación no supervisada (K-means) para clasificar los hogares en grupos homogéneos según su estructura de ingresos, gasto y ahorro, determinando el número óptimo de clusters mediante criterios estadísticos y asegurando su interpretación económica.
- **Evaluación de la capacidad predictiva del modelo:** desarrollar modelos supervisados (árboles de decisión y otros algoritmos) que permitan predecir el perfil financiero de los hogares a partir de variables estructurales, demográficas y territoriales, evaluando su rendimiento mediante métricas de clasificación.
- **Análisis de la aplicabilidad del modelo en el ámbito asegurador:** examinar el potencial del modelo como herramienta de segmentación avanzada, evaluando su utilidad para la identificación de perfiles de vulnerabilidad financiera y su posible integración en procesos de gestión del riesgo y diseño de productos en el sector asegurador.

1.3 ORGANIZACIÓN DE LA MEMORIA

El resto del trabajo se organiza de la siguiente manera. El capítulo 2 presenta la revisión de la literatura en torno a la medición de la pobreza, el concepto de salud financiera y la aplicación de técnicas de aprendizaje automático en análisis socioeconómico. El capítulo 3 describe la metodología del estudio de campo: fuente de datos (Encuesta de Presupuestos Familiares), variables, proceso de depuración y construcción del indicador de capacidad de ahorro. El capítulo 4 expone los resultados de la segmentación en perfiles, la caracterización de los clusters y el rendimiento del modelo predictivo (árbol de decisión). El capítulo 5 (Discusión) interpreta los hallazgos a la luz de los objetivos y la literatura, y señala limitaciones y líneas futuras. Las conclusiones recogen de forma sintética el cumplimiento de los objetivos, las implicaciones académicas y empresariales y las recomendaciones. Se incluyen así mismo referencias bibliográficas y anexos con material complementario.

Capítulo 2. MARCO TEÓRICO

Este capítulo revisa la literatura relevante para el diseño del modelo de salud financiera y, en particular, para la dimensión de capacidad de ahorro. Se abordan tres bloques: la medición tradicional de la pobreza y sus limitaciones; la evolución hacia el concepto de salud financiera y resiliencia económica; y la aplicación de técnicas de aprendizaje automático en el análisis socioeconómico y en la segmentación de hogares.

2.1 MEDICIÓN DE LA POBREZA Y SUS LIMITACIONES

La medición de la pobreza ha estado tradicionalmente vinculada al nivel de renta, constituyendo el enfoque dominante en la literatura económica durante décadas. Bajo esta perspectiva, la pobreza se define en función de umbrales absolutos o relativos de ingresos, siendo el indicador de riesgo de pobreza monetaria, inferior al 60% de la renta mediana, el más utilizado en el ámbito europeo.

Este enfoque presenta ventajas evidentes en términos de simplicidad, comparabilidad internacional y disponibilidad de datos. Sin embargo, resulta limitado cuando el objetivo es analizar la estabilidad financiera de los hogares o su capacidad para afrontar perturbaciones económicas. La renta constituye un flujo, pero no refleja necesariamente la estructura del gasto, el nivel de endeudamiento ni la capacidad de generar ahorro.

La literatura ha puesto de manifiesto que la pobreza es un fenómeno multidimensional que no puede ser capturado únicamente a través de indicadores monetarios. En este sentido, el enfoque de capacidades desarrollado por Sen (1985) amplía el análisis hacia las oportunidades reales de los individuos para alcanzar niveles de bienestar. Posteriormente, los índices de pobreza multidimensional han incorporado dimensiones adicionales como la educación, la salud o las condiciones de vida (Alkire y Foster, 2011)[3].

En el ámbito europeo, el indicador AROPE refleja esta evolución al integrar no solo la renta, sino también la privación material y la intensidad laboral. Este enfoque evidencia que una parte relevante de la población no considerada pobre en términos monetarios presenta vulnerabilidades en otras dimensiones.

Diversos estudios han señalado que dos hogares con niveles de renta similares pueden experimentar situaciones económicas radicalmente distintas dependiendo de su estructura presupuestaria (Atkinson, 1987)[4]. En este sentido, la pobreza monetaria puede infraestimar situaciones de vulnerabilidad financiera latente, especialmente en contextos caracterizados por elevados costes estructurales como la vivienda o la energía.

Por tanto, la literatura coincide en la necesidad de complementar los indicadores tradicionales con medidas que capturen el equilibrio presupuestario y la capacidad de adaptación de los hogares.

2.2 SALUD FINANCIERA Y RESILENCIA ECONÓMICA

patrones de comportamiento financiero que no son visibles mediante indicadores tradicionales. Por ejemplo, hogares con ingresos similares pueden pertenecer a perfiles completamente distintos en términos de estabilidad financiera.

Asimismo, la literatura conductual ha introducido nuevos elementos en el análisis del ahorro, como los sesgos cognitivos, la falta de educación financiera o la influencia del entorno institucional. Estos factores explican por qué hogares con características similares pueden presentar comportamientos financieros muy diferentes.

En este contexto, el uso de indicadores relativos como el ratio ahorro/ingresos permite analizar de forma más precisa el equilibrio presupuestario de los hogares. Este indicador elimina el efecto escala y facilita la comparabilidad entre distintos niveles de renta.

Además, el análisis del ahorro permite identificarComo evolución del enfoque tradicional, en los últimos años ha emergido el concepto de salud financiera, entendido como la capacidad de los hogares para cumplir con sus obligaciones económicas, absorber shocks y mantener estabilidad intertemporal sin recurrir de forma sistemática al endeudamiento.

Este concepto ha sido desarrollado por organismos como el Consumer Financial Protection Bureau (CFPB, 2015),[5] que define la salud financiera como la capacidad de gestionar el flujo de ingresos y gastos, afrontar imprevistos y planificar el futuro financiero.

A diferencia de los enfoques tradicionales, la salud financiera incorpora múltiples dimensiones: ingresos, gasto, endeudamiento, liquidez y capacidad de ahorro. En este marco, la estabilidad financiera se interpreta como un equilibrio dinámico entre ingresos y gastos.

La literatura ha destacado el papel central del ahorro como indicador de resiliencia financiera. La existencia de un colchón de ahorro permite a los hogares afrontar gastos imprevistos sin recurrir a mecanismos que deterioren su situación financiera futura, como el endeudamiento (Lusardi, 2015).[6] Por el contrario, la ausencia de ahorro o la presencia de déficits estructurales sitúan a los hogares en una posición de elevada vulnerabilidad.

No obstante, persisten dificultades en la medición empírica de la salud financiera, especialmente en la construcción de indicadores integrados que permitan operacionalizar este concepto de forma cuantitativa.

2.3 COMPORTAMIENTO DE AHORRO Y NIVEL DE INGRESOS

El comportamiento de ahorro ha sido ampliamente estudiado en la teoría económica, especialmente en el marco de la hipótesis del ciclo vital (Modigliani y Brumberg, 1954) y la hipótesis de la renta permanente (Friedman, 1957). Estas teorías sostienen que los individuos toman decisiones de consumo y ahorro en función de sus expectativas de ingresos a lo largo del tiempo, buscando suavizar su consumo.

Desde este enfoque, el ahorro no depende únicamente del nivel de ingresos actual, sino también de las expectativas futuras y del ciclo vital del individuo. En general, los hogares tienden a ahorrar más durante los años de mayor ingreso y a desahorrar en etapas como la jubilación.

No obstante, la evidencia empírica muestra que el comportamiento de ahorro está influido por múltiples factores adicionales. Entre ellos destacan el coste de la vivienda, el nivel de endeudamiento, la estructura del hogar, la edad o la incertidumbre económica.

En economías como la española, caracterizadas por elevados costes estructurales, el margen de ahorro puede verse limitado incluso en hogares con ingresos medios. Esto pone de manifiesto la importancia de analizar no solo el nivel de ingresos, sino también la estructura del gasto.

Asimismo, la literatura conductual ha introducido nuevos elementos en el análisis del ahorro, como los sesgos cognitivos, la falta de educación financiera o la influencia del entorno institucional. Estos factores explican por qué hogares con características similares pueden presentar comportamientos financieros muy diferentes.

En este contexto, el uso de indicadores relativos como el ratio ahorro/ingresos permite analizar de forma más precisa el equilibrio presupuestario de los hogares. Este indicador elimina el efecto escala y facilita la comparabilidad entre distintos niveles de renta.

Además, el análisis del ahorro permite identificar patrones de comportamiento financiero que no son visibles mediante indicadores tradicionales. Por ejemplo, hogares con ingresos similares pueden pertenecer a perfiles completamente distintos en términos de estabilidad financiera.

2.4 APLICACIÓN DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO EN ANÁLISIS SOCIOECONÓMICO

El desarrollo de técnicas de aprendizaje automático ha transformado el análisis de datos en múltiples disciplinas, incluida la economía. Estas herramientas permiten identificar patrones complejos y relaciones no lineales que no pueden ser capturados mediante métodos econométricos tradicionales (Hastie et al., 2009).

En el ámbito socioeconómico, el aprendizaje automático se utiliza cada vez más para segmentar poblaciones, predecir comportamientos y analizar grandes volúmenes de datos. Este enfoque resulta especialmente útil cuando se trabaja con microdatos, donde la heterogeneidad de las observaciones es elevada.

El aprendizaje automático se divide en dos grandes categorías: aprendizaje no supervisado y aprendizaje supervisado. El primero se utiliza para descubrir estructuras latentes en los datos sin etiquetas previas, mientras que el segundo permite predecir una variable objetivo conocida.

Dentro del aprendizaje no supervisado, los métodos de clustering permiten agrupar observaciones en función de su similitud. El algoritmo K-means es uno de los más utilizados debido a su simplicidad y eficiencia. Este algoritmo busca minimizar la variabilidad interna de los grupos, generando clusters homogéneos (Jain, 2010).

Por otro lado, los modelos supervisados, como los árboles de decisión, permiten construir modelos predictivos interpretables. Estos modelos generan reglas de clasificación basadas en variables observables, lo que facilita su aplicación en contextos empresariales.

La combinación de técnicas no supervisadas y supervisadas permite desarrollar un enfoque metodológico completo: en una primera fase se identifican perfiles de comportamiento, y en una segunda fase se construyen modelos que permiten predecir dichos perfiles.

Este enfoque resulta especialmente adecuado para el análisis de la salud financiera, ya que permite capturar la complejidad del comportamiento económico de los hogares y transformar esta información en herramientas útiles para la toma de decisiones.

En este sentido, el presente trabajo se sitúa en la intersección entre economía aplicada y ciencia de datos, aportando una aproximación innovadora al análisis de la vulnerabilidad financiera.

Capítulo 3. METODOLOGÍA

Este capítulo constituye el núcleo metodológico del trabajo. Se presenta de forma integrada la fundamentación teórica de cada decisión metodológica adoptada, el diseño del proceso de análisis y la descripción técnica de las herramientas empleadas. El objetivo es garantizar que el lector comprenda no solo qué se ha hecho, sino por qué cada elección metodológica responde a criterios teóricos y empíricos sólidos.

El enfoque adoptado combina técnicas de aprendizaje no supervisado y supervisado, articuladas en tres fases secuenciales: (i) construcción y validación de indicadores de salud financiera a partir de microdatos; (ii) identificación de perfiles homogéneos de comportamiento financiero mediante clustering; y (iii) desarrollo de modelos predictivos para anticipar la pertenencia a dichos perfiles. Esta estructura responde a la naturaleza del problema: la salud financiera de los hogares es un fenómeno latente, no directamente observable, que debe ser operacionalizado, segmentado y modelizado de forma supervisada para resultar útil en contextos aplicados.

3.1 MARCO TEÓRICO DE LAS TÉCNICAS METODOLÓGICAS EMPLEADAS

Esta sección expone en profundidad los fundamentos teóricos de cada técnica metodológica empleada en el trabajo. La justificación teórica de los métodos es esencial para garantizar la validez interna del análisis y para establecer con claridad las condiciones bajo las que los resultados son interpretables.

3.1.1 FUNDAMENTOS TEÓRICOS DEL APRENDIZAJE AUTOMÁTICO EN EL ANÁLISIS SOCIOECONÓMICO

El aprendizaje automático (machine learning) constituye una rama de la inteligencia artificial que desarrolla algoritmos capaces de aprender patrones a partir de los datos, sin que dichos patrones sean explícitamente programados por el analista (Mitchell, 1997; Hastie et al., 2009)[7]. A diferencia de los métodos econométricos tradicionales, que parten de un modelo paramétrico especificado a priori, los algoritmos de aprendizaje automático adoptan un enfoque inductivo: los patrones emergen de los datos mediante un proceso de optimización.

Esta diferencia tiene implicaciones metodológicas relevantes. Los métodos econométricos tradicionales son especialmente adecuados cuando el objetivo es la inferencia causal: permiten estimar el efecto de una variable sobre otra bajo supuestos identificativos explícitos y cuantificar la incertidumbre asociada. Sin embargo, cuando el objetivo es la predicción o la identificación de estructuras latentes en los datos, los métodos de aprendizaje automático presentan ventajas comparativas: son capaces de capturar relaciones no lineales, interacciones de orden elevado y estructuras complejas que los modelos lineales no pueden modelizar de forma flexible.

En el contexto de este trabajo, el objetivo no es estimar el efecto causal de una variable sobre la salud financiera, sino identificar perfiles homogéneos de comportamiento (fase de clustering) y construir un modelo capaz de predecirlos a partir de características observables (fase supervisada). Este objetivo es intrínsecamente predictivo y exploratorio, lo que justifica la adopción de un enfoque basado en aprendizaje automático.

El uso de técnicas de aprendizaje automático en el análisis socioeconómico ha experimentado un crecimiento significativo en los últimos años. Varian (2014)[8] y Athey e Imbens (2017)[9] han destacado el potencial de estos métodos para la economía aplicada, especialmente en contextos con grandes volúmenes de datos y estructuras de relaciones complejas. En particular, los métodos de clustering han sido aplicados para la segmentación

de hogares en distintos ámbitos: análisis de la pobreza (Nolan y Whelan, 2010)[10], segmentación de mercados financieros y análisis del comportamiento del consumidor.

3.1.2 APRENDIZAJE NO SUPERVISADO: EL CLUSTERING COMO TÉCNICA DE SEGMENTACIÓN

El aprendizaje no supervisado agrupa a un conjunto de métodos que tienen como objetivo descubrir estructuras latentes en los datos sin hacer uso de etiquetas o categorías predefinidas. A diferencia del aprendizaje supervisado, en el que el modelo aprende a partir de ejemplos etiquetados, el aprendizaje no supervisado opera exclusivamente sobre los datos de entrada, buscando organizar las observaciones en función de su similitud interna.

Dentro del aprendizaje no supervisado, los métodos de clustering (o agrupamiento) constituyen la familia de técnicas más ampliamente utilizada. El objetivo del clustering es particionar un conjunto de observaciones en grupos de forma que la similitud entre observaciones del mismo grupo sea máxima y la similitud entre grupos sea mínima. Esta dualidad —cohesión interna y separación entre grupos— es el principio fundamental que subyace a todos los métodos de clustering.

Formalmente, dado un conjunto de n observaciones $X = \{x_1, x_2, ..., x_n\}$, donde cada $x_i \in \mathbb{R}^d$ es un vector de d características, el problema del clustering consiste en encontrar una partición $C = \{C_1, C_2, ..., C_k\}$ del conjunto X en k grupos, de forma que se minimice alguna función de pérdida que capture la heterogeneidad interna de los clusters.

La justificación teórica del uso de clustering en este trabajo radica en la naturaleza del problema. La salud financiera de los hogares es un fenómeno continuo y multidimensional que puede presentar estructuras de agrupamiento natural: es plausible que existan grupos de hogares con comportamientos financieros cualitativamente distintos hogares estructuralmente en déficit frente a hogares con capacidad de ahorro estable, separados por fronteras relativamente nítidas en el espacio de variables. El clustering permite

identificar estas estructuras de forma no paramétrica, sin imponer a priori el número de grupos ni sus características.

3.1.3 EL ALGORITMO K-MEANS: FUNDAMENTOS, PROPIEDADES Y LIMITACIONES

El algoritmo K-means es uno de los métodos de clustering más ampliamente utilizados en la práctica, tanto por su simplicidad computacional como por su solidez teórica. [11]

El objetivo formal del algoritmo K-means consiste en encontrar una partición del conjunto de datos en k clusters que minimice la inercia intra-cluster (suma de cuadrados dentro del cluster, WCSS):

$$J = \sum_{k} \sum_{i \epsilon C_k} |X_i - \mu_k|^2 \quad (1)$$

donde μ_k representa el centroide del cluster C_k, definido como la media aritmética de las observaciones asignadas a dicho cluster. El algoritmo procede de forma iterativa mediante dos pasos alternados:

Paso de asignación: cada observación x_i se asigna al cluster cuyo centroide es más cercano en términos de distancia euclídea:

$$C_k = \left\{ x_i : |x_i - \mu_k|^2 \leq \left| x_i - \mu_j \right|^2, \forall j \neq k \right\} \quad (2)$$

Paso de actualización: los centroides se recalculan como la media de las observaciones asignadas a cada cluster:

$$\mu_k = (\frac{1}{|C_k|}) \sum_{i \in C_k} Xi \quad (3)$$

Este proceso se repite hasta que las asignaciones no cambian entre iteraciones, garantizando la convergencia del algoritmo. Es importante destacar que K-means converge a un óptimo local, no necesariamente al óptimo global. Para mitigar este problema, es práctica habitual ejecutar el algoritmo múltiples veces con inicializaciones aleatorias diferentes y seleccionar la solución con menor inercia intra-cluster.

Desde un punto de vista teórico, el algoritmo K-means presenta las siguientes propiedades relevantes para su aplicación en este trabajo:

- Asunción de esfericidad. K-means asume implícitamente que los clusters tienen forma aproximadamente esférica (covarianzas similares e isótropas). El algoritmo es relativamente robusto ante desviaciones moderadas de este supuesto, aunque puede generar resultados subóptimos cuando los clusters presentan formas muy irregulares o tamaños muy dispares.
- Sensibilidad a la escala. Dado que el algoritmo se basa en distancias euclídeas, es sensible a la escala de las variables. Variables con mayor magnitud o varianza contribuyen de forma desproporcionada al cálculo de distancias, lo que puede sesgar la segmentación. Esta propiedad justifica la necesidad de estandarizar las variables antes de aplicar el algoritmo, tal como se detalla en la sección de preprocesamiento.
- Sensibilidad a valores atípicos. La presencia de outliers puede distorsionar la posición de los centroides, afectando a la calidad de la segmentación. El proceso de depuración descrito anteriormente aborda esta limitación.
- Determinismo condicionado a la inicialización. Dado el mismo punto de partida, el algoritmo producirá siempre el mismo resultado. La variabilidad derivada de las distintas inicializaciones se gestiona mediante la ejecución múltiple con semillas aleatorias distintas.

Para el presente análisis se utilizaron como variables de entrada los ingresos anuales del hogar, el gasto anual total y el ratio ahorro/ingresos, todas ellas estandarizadas mediante z-score. La elección de estas tres variables responde al objetivo de capturar simultáneamente el nivel absoluto de ingresos y gasto, y el equilibrio relativo entre ambas magnitudes, de forma que hogares con niveles similares pero distinto equilibrio presupuestario pueda quedar asignados a clusters diferentes.

3.1.4 DETERMINACIÓN DEL NÚMERO ÓPTIMO DE CLUSTERS: EL MÉTODO DEL CODO

Una de las decisiones más relevantes en la aplicación del algoritmo K-means es la selección del número de clusters k. Esta elección no es trivial, ya que no existe un valor universalmente óptimo: la solución más adecuada depende del fenómeno estudiado, de los datos disponibles y del objetivo del análisis.

La literatura ha propuesto numerosos criterios para orientar esta decisión. En este trabajo se utiliza el método del codo (elbow method), que consiste en representar la inercia intra-cluster J(k) como función del número de clusters k y buscar el punto a partir del cual la reducción marginal de la inercia se vuelve decreciente de forma notable. La intuición subyacente es que, al aumentar k, la inercia disminuye necesariamente, pero esta disminución es decreciente: a partir de cierto punto, añadir más clusters mejora la inercia de forma marginal sin generar una ganancia de información económicamente significativa.

El "codo" de la curva representa el punto de equilibrio entre complejidad del modelo (mayor k) y ganancia de ajuste (menor inercia). Este criterio es intuitivo y transparente, lo que facilita su justificación en contextos aplicados. No obstante, presenta la limitación de que la identificación visual del codo puede ser ambigua cuando la curva es suave. Para complementar este análisis, pueden emplearse criterios adicionales como el índice de silueta, que mide la cohesión interna de los clusters y su separación respecto a los vecinos.

En este trabajo, el método del codo indica k = 3 como el número de clusters más adecuado. Esta elección se sustenta adicionalmente en consideraciones de interpretabilidad económica: tres perfiles (déficit estructural, ahorrador moderado y alta capacidad financiera)

permiten capturar la heterogeneidad relevante del comportamiento financiero sin incurrir en una sobresegmentación que dificulte la interpretación de los resultados.

3.1.5 APRENDIZAJE SUPERVISADO: FUNDAMENTOS TEÓRICOS DE LOS ÁRBOLES DE DECISIÓN

El aprendizaje supervisado engloba a los métodos que aprenden una función de predicción f: X → Y a partir de un conjunto de ejemplos etiquetados $\{(x_i, y_i)\}$, donde x_i es un vector de características y y_i es la etiqueta correspondiente. En el contexto de este trabajo, las etiquetas y_i corresponden a los perfiles financieros identificados en la fase de clustering, y el objetivo es aprender una función que prediga el perfil de un hogar a partir de sus características estructurales, demográficas y territoriales.

El modelo supervisado elegido es el árbol de decisión (decision tree), un método de clasificación que construye una estructura jerárquica de reglas binarias sobre las variables predictoras.

Un árbol de decisión divide recursivamente el espacio de características X en regiones rectangulares, asignando a cada región la clase mayoritaria entre los ejemplos de entrenamiento que caen en ella. El proceso de construcción del árbol se realiza mediante un algoritmo voraz que, en cada nodo, selecciona la variable y el umbral de corte que maximizan algún criterio de pureza de los subconjuntos resultantes.

Los criterios de pureza más comúnmente utilizados son:

- *Índice de Gini*: mide la probabilidad de clasificar incorrectamente una observación escogida aleatoriamente según la distribución de clases del nodo. Se define como:

$$G = \sum_k p_k(1 - p_k) \tag{4}$$

donde p_k es la proporción de observaciones de clase k en el nodo. Un nodo es puro (G = 0) cuando todas sus observaciones pertenecen a la misma clase.

- *Entropía*: mide la impureza del nodo en términos de información:

$$H = -\sum_{k} p_k \log_2 p_k \tag{5}$$

Un nodo puro tiene entropía cero. En este trabajo se utiliza el índice de Gini como criterio de división, en consonancia con la implementación estándar del algoritmo CART.

- *Error de clasificación*: proporción de observaciones mal clasificadas en el nodo. Menos sensible que el índice de Gini y la entropía para guiar el crecimiento del árbol.

3.2 FUENTE DE DATOS Y MUESTRA

3.2.1 ENCUESTA DE PRESUPUESTOS FAMILIARES (EPF): CARACTERIZACIÓN Y FUNDAMENTO TEÓRICO

El análisis empírico se basa en los microdatos de la Encuesta de Presupuestos Familiares (EPF), elaborada por el Instituto Nacional de Estadística (INE). Esta encuesta constituye la principal fuente estadística de ámbito nacional para el estudio del comportamiento económico de los hogares en España, proporcionando información detallada y representativa sobre ingresos, gastos y características sociodemográficas a escala de hogar.

Desde un punto de vista estadístico, la EPF es una operación continua que se realiza ininterrumpidamente desde el año 2006, con metodología armonizada a nivel europeo conforme al Reglamento CE 2454/97 y sus sucesivas modificaciones. La encuesta utiliza un diseño muestral bietápico estratificado: en la primera etapa se seleccionan secciones censales

con probabilidad proporcional al tamaño; en la segunda etapa se seleccionan viviendas dentro de cada sección. Este diseño garantiza la representatividad tanto a nivel nacional como a nivel de comunidad autónoma.

El uso de microdatos resulta especialmente adecuado para la aplicación de técnicas de aprendizaje automático por varias razones teóricas y prácticas. En primer lugar, los microdatos permiten trabajar a nivel individual de observación, preservando la heterogeneidad intrínseca de los hogares que se perdería al agregar la información. Esta heterogeneidad es precisamente el objeto de estudio en el presente trabajo, ya que el análisis pretende identificar perfiles diferenciados de comportamiento financiero. En segundo lugar, el volumen de observaciones disponibles satisface los requisitos mínimos de los algoritmos de clustering y clasificación supervisada, permitiendo obtener estimaciones estables y generalizables. En tercer lugar, la riqueza de variables recogidas en la EPF, ingresos, gastos desagregados por categorías, características del hogar, variables demográficas del sustentador principal y variables territoriales, facilita la construcción de indicadores multidimensionales de salud financiera.

Desde un punto de vista conceptual, la utilización de la EPF responde a las recomendaciones de organismos internacionales como Eurostat, el Fondo Monetario Internacional y la OCDE, que señalan las encuestas de presupuestos familiares como la fuente más adecuada para el análisis de la distribución del bienestar económico y la medición de la vulnerabilidad financiera a escala de hogar. En particular, la EPF ofrece una medida directa del gasto de consumo, que constituye la variable central para la construcción del indicador de capacidad de ahorro definido en este trabajo.

3.2.2 SELECCIÓN MUESTRAL Y PROCESO DE DEPURACIÓN

La muestra utilizada en el análisis empírico corresponde a la última edición disponible de la EPF. Antes de proceder al análisis, se aplicó un protocolo de depuración estadística con el objetivo de garantizar la consistencia interna de los datos y la robustez de

los indicadores construidos. El proceso incluyó las siguientes actuaciones, cada una justificada por criterios teóricos y técnicos:

- **Exclusión de hogares con ingresos nulos o negativos.** Los hogares que declaran ingresos anuales nulos o negativos no permiten la construcción del ratio ahorro/ingresos, ya que este indicador está definido como el cociente entre el excedente presupuestario y los ingresos totales. La inclusión de estas observaciones introduciría valores indefinidos o singularidades matemáticas que distorsionarían los resultados. Desde un punto de vista económico, los ingresos nulos pueden reflejar situaciones atípicas, hogares en proceso de liquidación, errores de respuesta o casos de autoconsumo no monetario, que no son representativas del comportamiento financiero de los hogares en sentido convencional.
- **Tratamiento de valores perdidos.** Los valores perdidos en las variables clave se identificaron y trataron mediante análisis de patrones de ausencia. Cuando la proporción de valores perdidos era reducida y no sistemática, se optó por la exclusión listwise de las observaciones afectadas, dado que el volumen muestral lo permite sin comprometer la representatividad. Esta decisión es preferible a la imputación cuando el porcentaje de datos ausentes es bajo, ya que evita introducir sesgo derivado del método de imputación.
- **Acotación del ratio ahorro/ingresos al intervalo [−1, 1].** La construcción del ratio puede generar valores extremos en hogares con ingresos muy bajos y gastos muy elevados. Para garantizar la estabilidad del indicador, se aplicó un proceso que limita los valores del ratio al intervalo mencionado. Esta decisión reduce la influencia de observaciones atípicas en los algoritmos de clustering, que son especialmente sensibles a la presencia de outliers.
- **Verificación de coherencia interna.** Se comprobó la coherencia entre las variables de ingresos y gasto declaradas, identificando y excluyendo observaciones con inconsistencias manifiestas que podrían indicar errores de codificación en la encuesta original.

3.3 CONSTRUCCIÓN DE VARIABLES

El núcleo del modelo propuesto se basa en la construcción de indicadores que capturen distintas dimensiones de la salud financiera de los hogares. Entre ellos, destaca la variable de capacidad de ahorro, definida como:

- Diferencia entre ingresos y gastos
- Normalizada mediante el ratio ahorro/ingresos

La capacidad de ahorro se definió como el ratio:

$$ratio\,\frac{ahorro}{ingresos} = \frac{ingresos - gasto}{ingresos} \tag{6}$$

donde ingresos representa los ingresos anuales totales del hogar y gasto el gasto anual total. Este indicador mide el excedente presupuestario (o el déficit, si es negativo) en términos relativos, permitiendo comparar hogares de distinto tamaño económico.

Este enfoque permite eliminar el efecto escala y comparar hogares con distintos niveles de renta, facilitando la identificación de patrones estructurales.

Con el fin de garantizar la robustez del indicador, se ha aplicado un proceso de transformación que limita el ratio al intervalo [-1,1], evitando la influencia de valores extremos que puedan distorsionar los resultados. Esta decisión responde a la necesidad de estabilizar la distribución de la variable y mejorar el rendimiento de los algoritmos de segmentación.

Adicionalmente, se han construido variables complementarias relacionadas con la estructura del gasto, el esfuerzo económico y las características sociodemográficas del hogar, con el objetivo de enriquecer el análisis y capturar la multidimensionalidad del fenómeno estudiado.

3.3.1 DISTRIBUCIÓN DEL RATIO Y NIVEL DE AHORRO

La figura muestra la distribución del ratio ahorro/ingresos tras el proceso de depuración y eliminación de valores extremos. La mayor parte de los hogares presentan valores positivos del indicador, concentrándose entre 0 y 0,6, lo que sugiere la existencia de superávit presupuestario en una proporción significativa de la muestra. No obstante, se identifica un subconjunto de hogares con valores negativos, reflejando situaciones de déficit estructural. La asimetría de la distribución confirma la heterogeneidad en el comportamiento financiero y justifica la aplicación de técnicas de segmentación para identificar perfiles homogéneos en lugar de limitarse a estadísticos descriptivos globales.

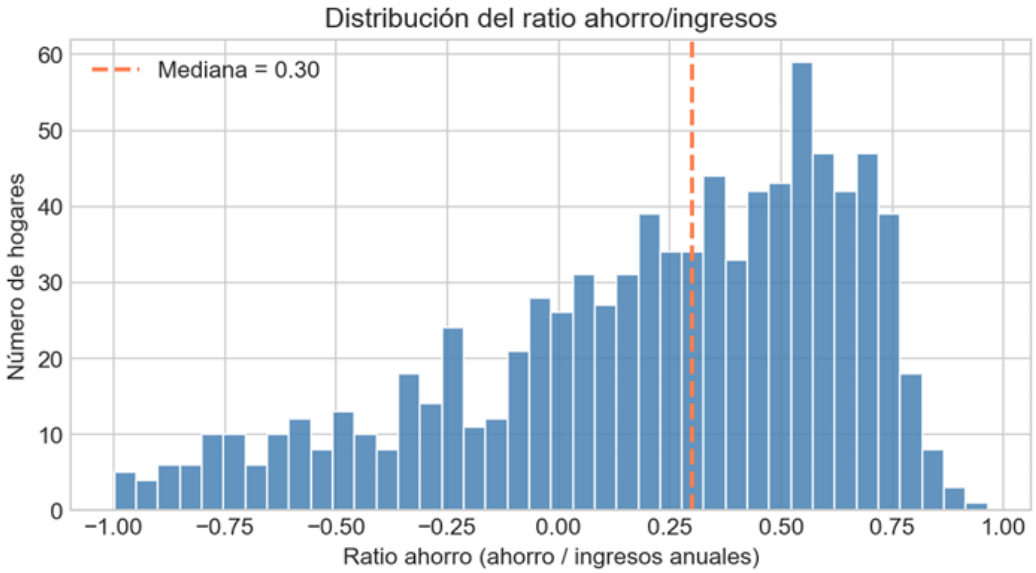


Figura 1 - Distribución del ratio ahorro/ingresos tras el proceso de depuración

A partir del ratio se construyó una variable categórica de nivel de ahorro para describir la muestra. Las figuras 2 y 3 muestran la distribución de hogares por nivel de ahorro (barras) y el boxplot del ratio por nivel, respectivamente.

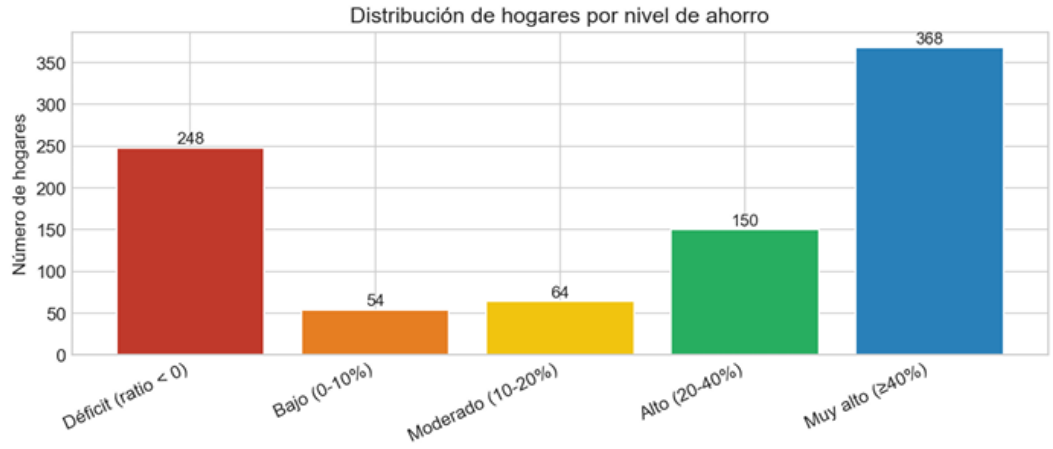


Figura 2 - Distribución por nivel de ahorro (barras)

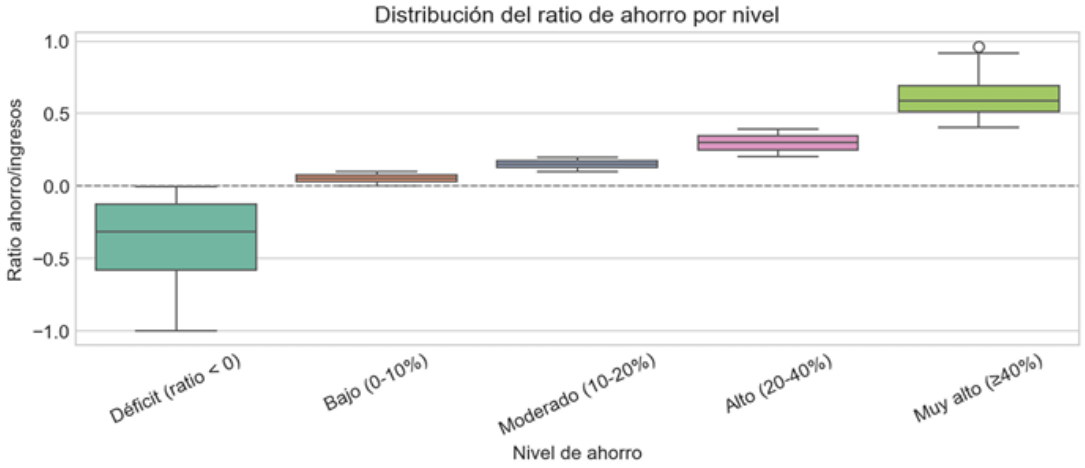


Figura 3 - Boxplot del ratio ahorro/ingresos por nivel de ahorro

En términos de ahorro/ingresos, estos niveles permiten interpretar de forma directa la posición presupuestaria de cada hogar. Los hogares en déficit son aquellos cuyo gasto anual supera a sus ingresos, de modo que el ratio es negativo: viven por encima de sus posibilidades presentes y dependen de endeudamiento o de desahorro de patrimonio acumulado. Los hogares en equilibrio presentan un ratio próximo a cero: sus ingresos cubren prácticamente sus gastos, pero disponen de un margen muy reducido para acumular ahorro.

Finalmente, los hogares con ahorro medio o alto muestran ratios crecientes, lo que indica una capacidad de generación de excedente que, en términos de salud financiera, se asocia con una mayor resiliencia ante shocks.

3.4 SEGMENTACIÓN DE HOGARES MEDIANTE CLUSTERING

Para identificar perfiles homogéneos de comportamiento financiero, se ha aplicado el algoritmo de clustering K-means, ampliamente utilizado en el análisis exploratorio de datos.

El objetivo de este algoritmo es particionar la muestra en $k$grupos de forma que se minimice la variabilidad interna de cada cluster y se maximice la distancia entre clusters. En este contexto, cada grupo representa un perfil financiero caracterizado por una combinación específica de ingresos, gasto y capacidad de ahorro.

La selección del número óptimo de clusters se ha realizado mediante criterios estadísticos como el método del codo y el análisis de la varianza explicada. Este paso resulta fundamental para evitar tanto la sobre segmentación como la pérdida de información relevante.

Los clusters obtenidos permiten identificar distintos perfiles, tales como:

- Hogares con déficit estructural
- Hogares con equilibrio presupuestario
- Hogares con alta capacidad de ahorro

Esta segmentación constituye la base para el análisis posterior y permite transformar variables continuas en categorías interpretables desde un punto de vista económico.

3.5 EVALUACIÓN DEL MODELO

El rendimiento del modelo predictivo se evalúa mediante un conjunto de métricas estándar de clasificación multiclase:

- Exactitud global (Accuracy). Proporción de observaciones correctamente clasificadas sobre el total del conjunto de prueba. Es la métrica más intuitiva, pero puede resultar engañosa en presencia de clases desequilibradas.
- Matriz de confusión. Tabla que cruza las clases reales (filas) con las clases predichas (columnas). Los elementos de la diagonal representan las clasificaciones correctas, mientras que los elementos fuera de la diagonal recogen los errores de predicción. La lectura de la matriz permite identificar en qué perfiles el modelo acierta más o menos, y si existen confusiones sistemáticas entre perfiles adyacentes (por ejemplo, entre ahorrador moderado y alta capacidad financiera).
- Precisión (Precision). Para cada clase k, proporción de observaciones predichas como clase k que efectivamente pertenecen a dicha clase. Mide la fiabilidad de las predicciones positivas del modelo.
- Sensibilidad o exhaustividad (Recall). Para cada clase k, proporción de observaciones que realmente pertenecen a la clase k y que el modelo clasifica correctamente. Mide la capacidad del modelo para detectar todas las observaciones de cada clase.
- Puntuación F1 (F1-score). Media armónica de la precisión y la sensibilidad, que combina ambas métricas en un único indicador. Es especialmente útil cuando existe desequilibrio entre clases.

Además de estas métricas cuantitativas, se evalúa la coherencia económica de los resultados, verificando que los perfiles identificados y las reglas de clasificación sean interpretables y consistentes con la teoría económica revisada en el marco teórico. Un modelo con alta exactitud técnica pero resultados económicamente incoherentes no resulta útil para los fines aplicados del presente trabajo.

Capítulo 4. ANÁLISIS DE RESULTADOS

En este capítulo se presentan los resultados del análisis correspondiente a la dimensión de capacidad de ahorro: la segmentación de hogares en perfiles, su caracterización y el rendimiento del modelo predictivo. Los resultados se organizan en tres secciones: segmentación en perfiles de ahorro, caracterización de los perfiles y número óptimo de clusters y modelo predictivo

4.1 SEGMENTACIÓN EN PERFILES DE AHORRO

Se aplicó el algoritmo K-means sobre las variables ingresos anuales, gasto anual y ratio ahorro/ingresos (estandarizadas) para identificar perfiles homogéneos de comportamiento presupuestario. La elección de estas tres variables permite capturar no solo el nivel absoluto de ingresos y gasto, sino la relación entre ambos a través del ratio, de modo que hogares con niveles similares pero distinto equilibrio presupuestario pueda quedar en clusters diferentes. La figura 4 muestra la proyección de los hogares en el espacio de las variables consideradas y la asignación a cada perfil mediante colores o símbolos diferenciados.

Se identifican tres perfiles principales, que se han etiquetado de forma interpretable:

- *Déficit Estructural*: hogares cuyo gasto supera de forma sistemática a los ingresos (ratio negativo o muy bajo), lo que indica dependencia del desahorro o del endeudamiento.
- *Ahorrador Moderado*: hogares en situación de equilibrio o con un superávit moderado (ratio positivo pero no extremo), que pueden mantener un cierto margen de ahorro sin destacar por una capacidad de ahorro muy elevada.
- *Alta Capacidad Financiera*: hogares con un ratio de ahorro elevado, que destinan una proporción importante de sus ingresos al ahorro o a la reducción de deuda. Esta

triple clasificación resulta coherente con la literatura revisada y con el objetivo de disponer de perfiles interpretables para la segmentación empresarial y el análisis de riesgo.

Desde la perspectiva específica de la relación ahorro/ingresos, los tres perfiles permiten responder a preguntas clave: qué parte de la población se encuentra atrapada en un déficit crónico, con un ratio negativo que apenas mejora con aumentos marginales de ingresos; qué proporción de hogares se sitúa en una zona intermedia de equilibrio frágil, donde pequeños shocks pueden hacerles pasar de un ligero superávit a un déficit; y qué grupos de población consolidan una capacidad de ahorro alto y estable. El análisis de los centroides y de los gráficos posteriores permite vincular estas situaciones con tramos de renta, niveles de gasto y, en la medida en que se incorporen variables demográficas, con características como el tamaño del hogar o la edad del sustentador principal.

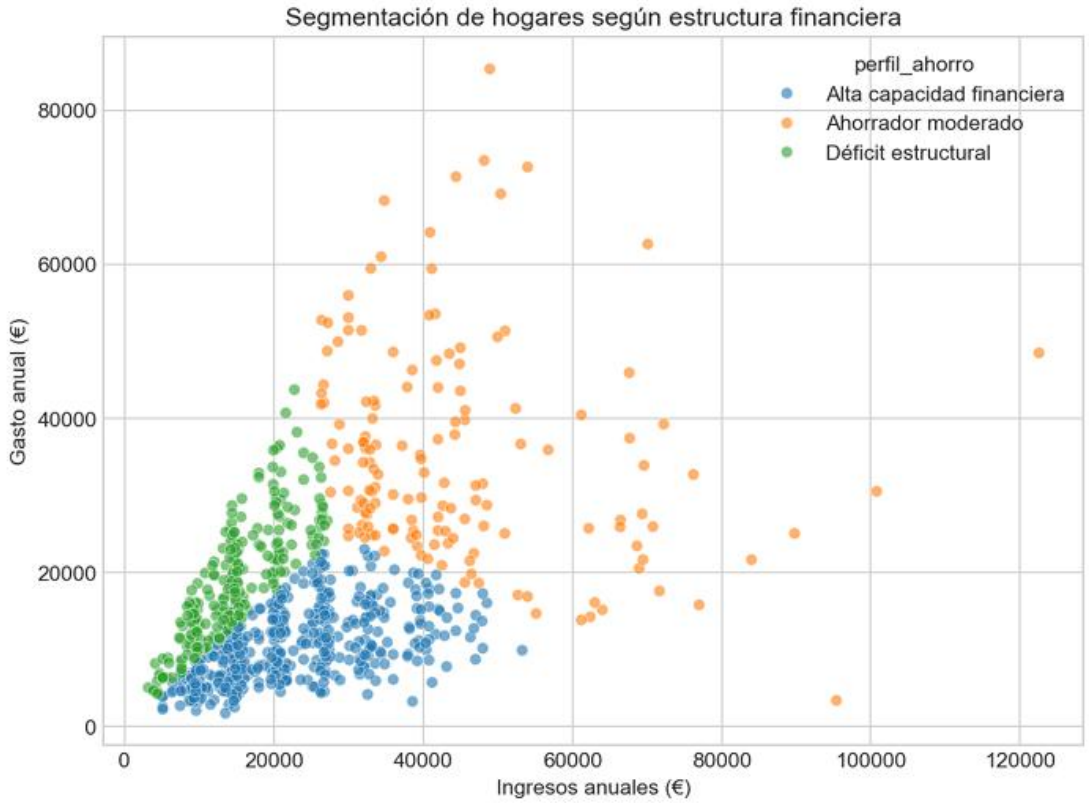


Figura 4 - Segmentación de hogares por perfil de ahorro (K-means)

La figura 5 representa la relación entre ingresos y gasto en el conjunto de la muestra. Los hogares situados por encima de la bisectriz (línea en la que ingresos = gasto) presentan superávit; los situados por debajo, déficit. La nube de puntos permite visualizar la dispersión de los hogares y la concentración de cada perfil en distintas zonas del plano. La segmentación obtenida con K-means se refleja en la agrupación espacial: los hogares en déficit estructural se sitúan predominantemente por debajo de la bisectriz, mientras que los de alta capacidad financiera se concentran en la zona de ingresos altos y gasto relativamente menor.

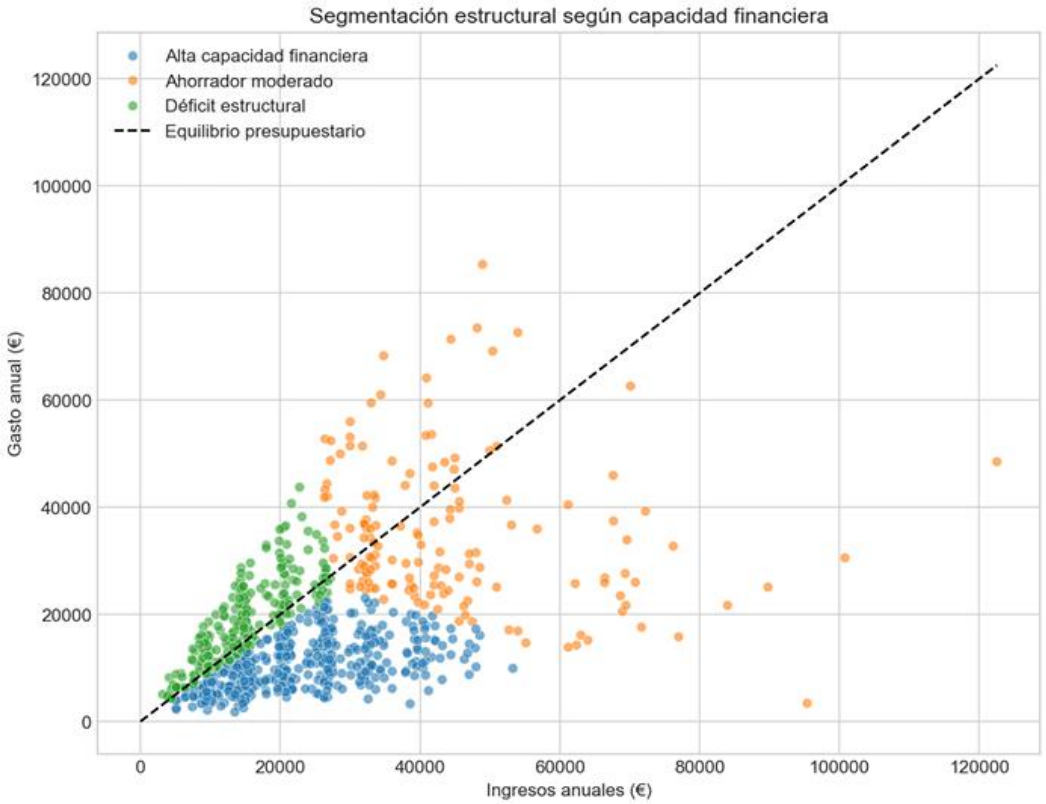


Figura 5 - Ingresos anuales frente a gastos anual: equilibrio presupuestario

Los resultados muestran que una proporción significativa de los hogares presenta superávit presupuestario anual. Este resultado no implica necesariamente ausencia de vulnerabilidad financiera, sino que refleja que, en términos contables, el gasto declarado es inferior a los ingresos anuales en una parte relevante de la muestra.

Sin embargo, el análisis también identifica un grupo estructural de hogares cuyo gasto supera sistemáticamente a sus ingresos, configurando un perfil de déficit persistente. Este grupo representa el núcleo de mayor fragilidad financiera dentro de la dimensión analizada.

Además, incluso entre los hogares con superávit, se observan diferencias significativas en la magnitud del excedente y en el nivel absoluto de ingresos, lo que evidencia una heterogeneidad relevante en la estabilidad financiera.

En consecuencia, la segmentación permite distinguir entre:

- Superávit elevado con alta capacidad financiera.
- Superávit moderado en hogares de renta media.
- Déficit estructural persistente.

Esta diferenciación aporta información más rica que el análisis unidimensional basado exclusivamente en renta.

4.2 CARACTERIZACIÓN DE LOS PERFILES

Para validar la coherencia de la segmentación y facilitar su interpretación económica, se caracterizaron los tres perfiles mediante sus centroides, la distribución de frecuencias y las medias de las variables clave. Los centroides de cada cluster representan el "centro" del grupo en el espacio de variables (ingresos, gasto, ratio) y permiten comparar los perfiles entre sí. La figura 6 muestra la posición de los centroides en el espacio de variables; las figuras 7 y 10 muestran la distribución de hogares por perfil en forma de barras y de diagrama de sectores (pie), respectivamente; la figura 8 presenta las medias de ingresos, gasto y ratio ahorro/ingresos por perfil; la figura 9 muestra el boxplot del ratio por perfil, ilustrando la dispersión y los cuartiles dentro de cada grupo; y la figura 11 representa el diagrama de

dispersión de ingresos frente a ahorro (en niveles), coloreado por perfil, lo que permite visualizar la relación entre nivel de renta y capacidad de ahorro en cada grupo.

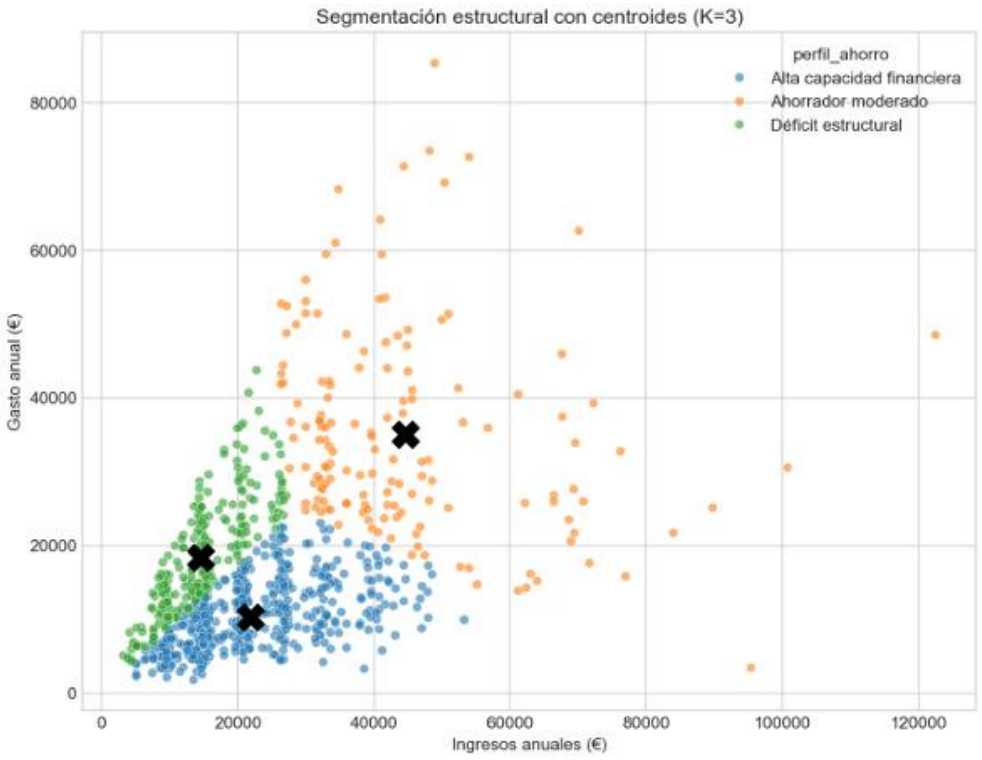


Figura 6 - Segmentación y centroides de los perfiles de ahorro

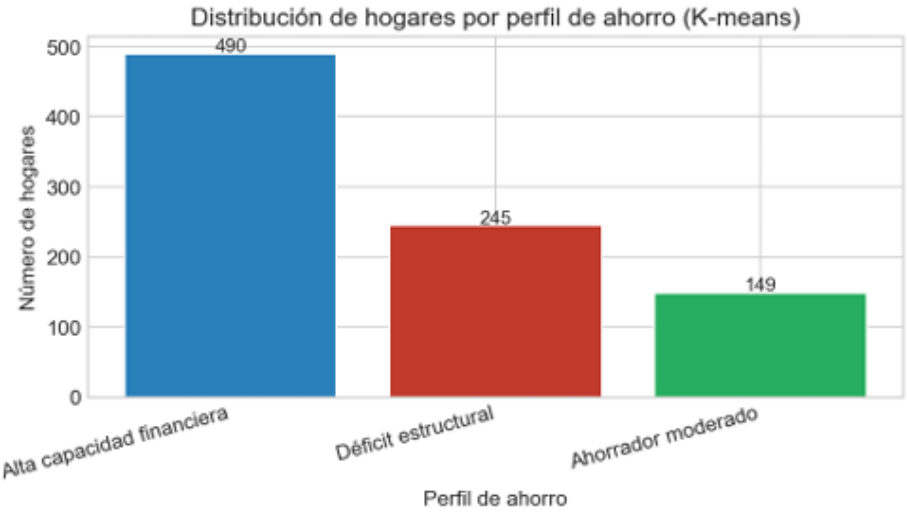


Figura 7 - Distribución de hogares por perfil (barras)



Figura 8 - Medias de ingresos, gasto y ratio ahorro/ingresos por perfil

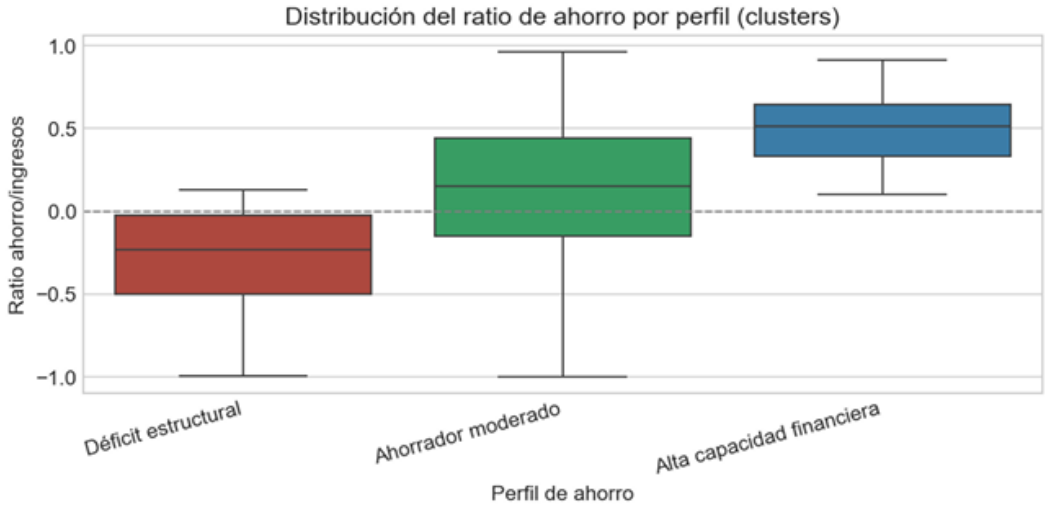


Figura 9 - Boxplot del ratio ahorro/ingresos por perfil



Figura 10 - Proporción de hogares por perfil (diagrama de sectores)

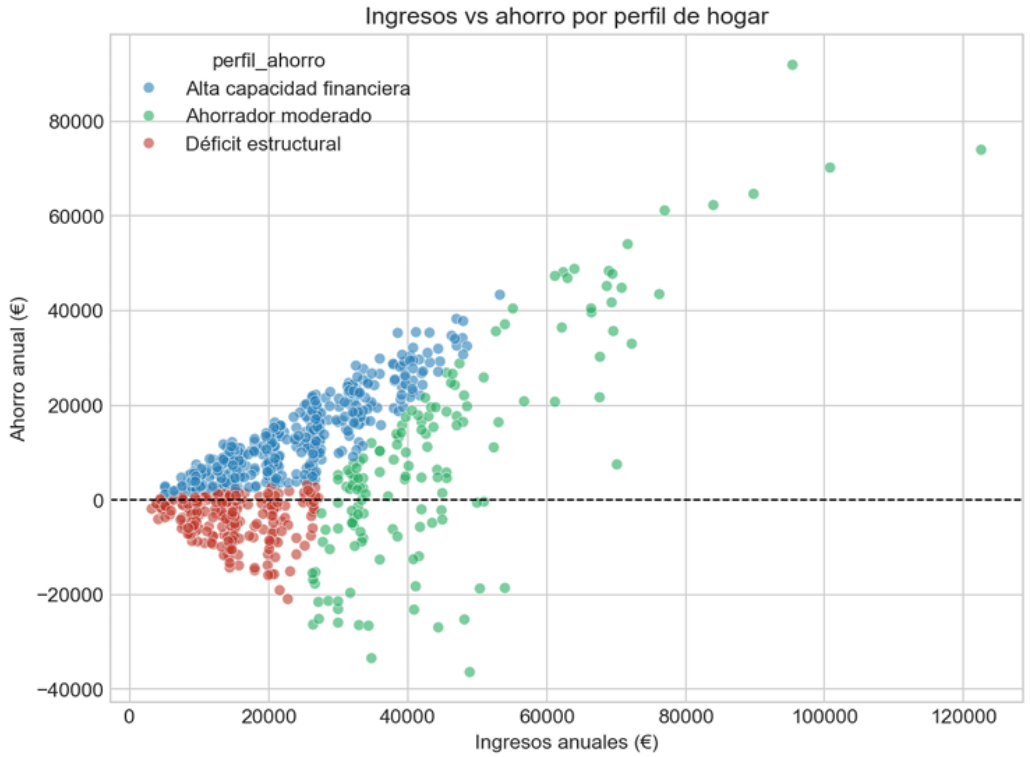


Figura 11 - Ingresos anuales frente a ahorro, coloreado por perfil

Los resultados confirman una heterogeneidad relevante entre los tres perfiles. El perfil de déficit estructural concentra hogares con ratio negativo o muy bajo; sus medias de ingresos y gasto suelen mostrar un gasto superior a los ingresos, coherente con la etiqueta asignada. El perfil ahorrador moderado agrupa a hogares con equilibrio o superávit moderado (ratio positivo pero no extremo), a menudo con niveles de ingresos y gasto intermedios. El perfil de alta capacidad financiera corresponde a hogares con ratio de ahorro elevado; en muchos casos presentan ingresos más altos y/o una estructura de gasto que deja un margen considerable para el ahorro. Las medias por perfil (figura 8) permiten interpretar económicamente cada cluster y validar que la segmentación no solo es estadísticamente coherente, sino también significativa desde el punto de vista de la salud financiera. La distribución en pie (figura 10) informa sobre el peso relativo de cada perfil en la muestra, lo cual es relevante para priorizar actuaciones en política social o en estrategia comercial.

4.3 NÚMERO ÓPTIMO DE CLUSTERS Y MODELO PREDICTIVO

Para determinar el número de clusters se utilizó el método del codo, que consiste en representar la evolución de la inercia intra-cluster (suma de distancias al cuadrado de cada observación a su centroide) en función del número k de clusters. Cuando k aumenta, la inercia disminuye; el "codo" de la curva indica un punto a partir del cual añadir más clusters no mejora sustancialmente la homogeneidad interna. La figura 12 muestra esta evolución; el codo sugiere un número de clusters coherente con la elección de tres perfiles, que además resultan interpretables y alineados con la literatura sobre perfiles de ahorro (déficit, equilibrio, superávit elevado). En ejercicios de robustez podría contrastarse con el índice de silueta u otros criterios.

Se entrenó un árbol de decisión para predecir el perfil de ahorro a partir de variables estructurales, demográficas y territoriales del hogar. El árbol de decisión ofrece reglas de clasificación en forma de umbrales sobre las variables (por ejemplo, si los ingresos son inferiores a un cierto valor y el gasto superior a otro, se asigna el perfil de déficit), lo que facilita su interpretación y su uso en procesos de scoring o segmentación.

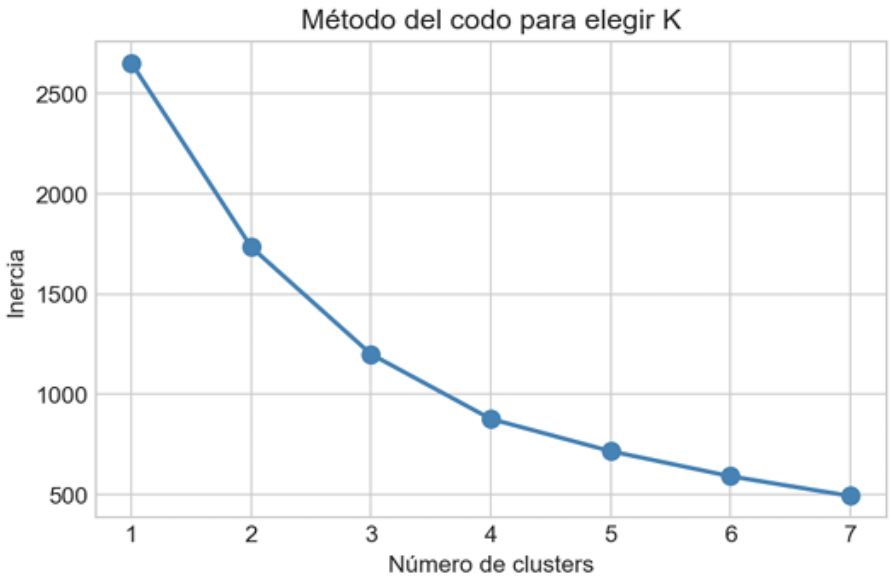


Figura 12 - Método del codo para elegir el número de clusters (K-means)

La figura 13 presenta la matriz de confusión del modelo sobre el conjunto de prueba: en las filas se encuentra la clase real (perfil observado) y en las columnas la clase predicha; los elementos de la diagonal representan las clasificaciones correctas. La lectura de la matriz permite evaluar en qué perfiles el modelo acierta más o menos y si existen confusiones sistemáticas (por ejemplo, entre ahorrador moderado y alta capacidad).

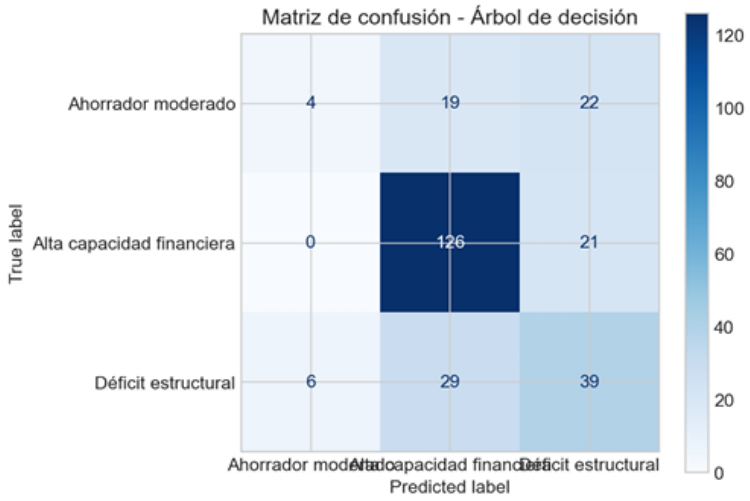


Figura 13 - Matriz de confusión del modelo de árbol de decisión (prefil de ahorro)

En conjunto, los resultados muestran que la dimensión de capacidad de ahorro permite segmentar de forma coherente a los hogares españoles, que los perfiles son interpretables desde un punto de vista económico y que un modelo supervisado como el árbol de decisión ofrece capacidad predictiva e interpretabilidad útiles para su integración en un score global de salud financiera y para aplicaciones en el ámbito asegurador. En el siguiente capítulo se discuten las implicaciones de estos hallazgos y se enmarcan en las limitaciones y líneas futuras del proyecto.

Capítulo 5. DISCUSIÓN

Este capítulo discute los resultados obtenidos a la luz de los objetivos planteados, la literatura revisada y sus implicaciones tanto académicas como aplicadas. Asimismo, se analizan las principales limitaciones del estudio y se proponen líneas futuras de investigación que permitan ampliar y mejorar el modelo desarrollado.

5.1 RELACIÓN CON LOS OBJETIVOS Y LA LITERATURA

El trabajo ha permitido cumplir los objetivos planteados en relación con la dimensión de capacidad de ahorro dentro del modelo de salud financiera. En particular, la construcción del ratio (ingresos − gasto)/ingresos y su depuración metodológica se alinean con la literatura que destaca la importancia del equilibrio presupuestario como indicador clave de resiliencia financiera (Lusardi, 2015)[6]. Este indicador permite capturar de forma directa la capacidad del hogar para generar excedente, superando las limitaciones de los enfoques basados exclusivamente en la renta.

Los resultados empíricos obtenidos refuerzan la idea, ampliamente discutida en la literatura, de que la vulnerabilidad financiera no puede entenderse como un fenómeno binario, sino como un continuum de situaciones. En este sentido, la identificación de tres perfiles diferenciados (déficit estructural, ahorrador moderado y alta capacidad financiera) es coherente con enfoques multidimensionales de análisis del bienestar económico, como los propuestos en estudios sobre pobreza relativa y privación material. La segmentación obtenida confirma que hogares con niveles de renta similares pueden presentar comportamientos financieros significativamente distintos en función de su estructura de gasto, lo que respalda la crítica a los indicadores tradicionales de pobreza monetaria (Atkinson, 1987)[4].

Desde el punto de vista metodológico, la aplicación del algoritmo K-means se sitúa en línea con la práctica habitual en el análisis socioeconómico (Jain, 2010),[12] donde las

técnicas de clustering permiten identificar estructuras latentes en los datos sin imponer categorías a priori. El uso del método del codo como criterio de selección del número de clusters refuerza la robustez del análisis, evitando problemas de sobre segmentación o pérdida de información relevante. Asimismo, la interpretación económica de los clusters obtenidos constituye un elemento clave, ya que permite traducir los resultados estadísticos en perfiles comprensibles y útiles para la toma de decisiones.

Por otro lado, la utilización de un modelo supervisado basado en árboles de decisión introduce una dimensión predictiva en el análisis. Este enfoque es coherente con la creciente incorporación de técnicas de aprendizaje automático en el ámbito económico, especialmente en contextos donde la interpretabilidad del modelo es un requisito fundamental. A diferencia de modelos más complejos como redes neuronales o métodos ensemble, los árboles de decisión permiten generar reglas claras de clasificación, lo que facilita su aplicación en entornos empresariales y regulatorios.

En términos más amplios, los resultados del trabajo se sitúan en la línea de la literatura reciente sobre salud financiera, que enfatiza la necesidad de integrar múltiples dimensiones en el análisis de la vulnerabilidad económica. Estudios previos han señalado la importancia de variables como el endeudamiento, la liquidez o la estabilidad del empleo, pero la evidencia empírica muestra que la capacidad de ahorro constituye un indicador especialmente relevante, ya que sintetiza el equilibrio entre ingresos y gastos. En este sentido, el trabajo contribuye a la literatura al ofrecer una operacionalización concreta y empíricamente validada de esta dimensión.

Además, los resultados obtenidos presentan implicaciones relevantes en el ámbito de la segmentación económica. La identificación de un grupo de hogares en déficit estructural persistente coincide con la literatura sobre exclusión financiera, que señala la existencia de colectivos atrapados en dinámicas de vulnerabilidad. Al mismo tiempo, la existencia de un grupo de hogares con superávit moderado sugiere la presencia de situaciones de equilibrio frágil, donde pequeños shocks económicos pueden provocar transiciones hacia estados de mayor vulnerabilidad. Este resultado es consistente con enfoques dinámicos de la pobreza que destacan la importancia de la movilidad entre estados financieros.

En conjunto, la evidencia empírica obtenida refuerza la necesidad de avanzar hacia modelos multidimensionales de análisis de la salud financiera, capaces de capturar la complejidad del comportamiento económico de los hogares y de superar las limitaciones de los indicadores tradicionales.

5.2 IMPLICACIONES Y LIMITACIONES

Los resultados del trabajo tienen implicaciones relevantes tanto desde una perspectiva académica como aplicada. En el ámbito académico, el estudio contribuye al debate sobre la medición de la vulnerabilidad financiera, proporcionando evidencia empírica sobre la utilidad de indicadores basados en el equilibrio presupuestario. En particular, la construcción del ratio ahorro/ingresos y su utilización en procesos de segmentación y predicción refuerzan la idea de que la salud financiera debe analizarse desde una perspectiva estructural y dinámica.

Desde una perspectiva empresarial, el modelo desarrollado presenta un alto potencial de aplicación en el sector asegurador y financiero. La identificación de perfiles homogéneos de comportamiento financiero permite mejorar la segmentación de clientes, facilitando el diseño de productos adaptados a distintos niveles de riesgo. Asimismo, la capacidad predictiva del modelo basado en árboles de decisión permite anticipar el comportamiento financiero de los hogares a partir de variables observables, lo que puede resultar especialmente útil en procesos de scoring, gestión del riesgo y toma de decisiones comerciales.

No obstante, los resultados deben interpretarse teniendo en cuenta una serie de limitaciones. En primer lugar, la Encuesta de Presupuestos Familiares (EPF) es una fuente basada en flujos de ingresos y gastos, lo que implica que no captura de forma completa la situación patrimonial de los hogares. La ausencia de información detallada sobre activos y pasivos limita la capacidad del indicador de ahorro para reflejar la posición financiera neta, especialmente en hogares con elevados niveles de endeudamiento o patrimonio acumulado.

En segundo lugar, la naturaleza transversal de los datos impide analizar la evolución temporal de los perfiles identificados. La estabilidad de los clusters y la movilidad de los hogares entre perfiles constituyen aspectos clave para comprender la dinámica de la vulnerabilidad financiera, por lo que su análisis requeriría el uso de datos longitudinales o la combinación de varias ediciones de la encuesta.

En tercer lugar, las decisiones metodológicas adoptadas en el proceso de depuración, como la exclusión de hogares con ingresos nulos o la acotación del ratio al intervalo $[-1,1]$, pueden afectar a la representatividad de los resultados en los extremos de la distribución. Si bien estas decisiones están justificadas desde el punto de vista técnico, es importante considerar su impacto potencial en la interpretación de los resultados.

Asimismo, el modelo predictivo se ha basado en árboles de decisión, priorizando la interpretabilidad frente a la precisión. La comparación con otros algoritmos más complejos, como random forests o redes neuronales, podría mejorar el rendimiento predictivo, aunque a costa de una menor transparencia en las reglas de clasificación.

Finalmente, es importante destacar que la dimensión de capacidad de ahorro constituye solo una parte del modelo global de salud financiera. La integración con otras dimensiones relevantes, como el esfuerzo en vivienda, el gasto energético o la privación material, resulta fundamental para construir un indicador sintético más completo. En este sentido, el trabajo sienta las bases para el desarrollo de un modelo multidimensional más amplio, que permita capturar de forma integral la vulnerabilidad financiera de los hogares.

Como líneas futuras de investigación, se propone la incorporación de información sobre endeudamiento y patrimonio, el análisis longitudinal de los perfiles financieros y la exploración de metodologías alternativas de aprendizaje automático. Asimismo, la aplicación del modelo en contextos reales, como la segmentación de clientes en el sector asegurador, permitiría validar su utilidad práctica y mejorar su capacidad predictiva.

A pesar de estas limitaciones, el trabajo proporciona una base metodológica sólida y resultados interpretables, que contribuyen tanto al avance del conocimiento académico como al desarrollo de herramientas aplicadas en el análisis de la salud financiera.

Capítulo 6. CONCLUSIONES

El presente trabajo ha permitido avanzar en los tres objetivos planteados para la dimensión de capacidad de ahorro dentro del modelo global de salud financiera de los hogares españoles, en el marco del proyecto desarrollado en colaboración con MAPFRE.

En primer lugar, se ha construido un indicador estructural de capacidad de ahorro, el ratio (ingresos − gasto)/ingresos, a partir de la Encuesta de Presupuestos Familiares, con un proceso de depuración (eliminación de ingresos nulos, tratamiento de valores faltantes) y acotación del ratio al intervalo [−1,1] para evitar distorsiones por valores extremos. La distribución del ratio refleja una heterogeneidad relevante: una parte significativa de los hogares presenta superávit presupuestario, con valores del ratio concentrados en el rango positivo; no obstante, existe un subconjunto de hogares con ratio negativo o muy bajo que refleja situaciones de déficit estructural. Estos hogares no son necesariamente identificados como pobres por los indicadores monetarios tradicionales, lo que refuerza la necesidad de complementar la medición de la pobreza con indicadores de equilibrio presupuestario y capacidad de ahorro. En segundo lugar, la segmentación mediante K-means ha permitido identificar tres perfiles homogéneos e interpretables: déficit estructural, ahorrador moderado y alta capacidad financiera. La caracterización por centroides, medias de ingresos y gasto y ratio, y la representación gráfica (barras, boxplots, diagramas de dispersión) validan la coherencia de los clusters y su utilidad para distinguir comportamientos presupuestarios claramente diferenciados. El método del codo respalda la elección de tres clusters como equilibrio entre homogeneidad interna y parsimonia. Los perfiles obtenidos son directamente utilizables para la segmentación de la población en estudios de riesgo, diseño de productos o análisis de políticas públicas.

En tercer lugar, el modelo predictivo (árbol de decisión) muestra que es posible anticipar el perfil de ahorro a partir de variables estructurales, demográficas y territoriales del hogar. La matriz de confusión permite evaluar el rendimiento del modelo en el conjunto de prueba, y el árbol de decisión ofrece reglas de clasificación interpretables que pueden

integrarse en procesos de scoring o de segmentación. La combinación de segmentación no supervisada y modelización supervisada aporta así no solo una descripción de los perfiles existentes, sino también una herramienta para predecir el perfil cuando no se dispone de información completa sobre ingresos y gasto, o para validar la coherencia de la información declarada.

Desde el punto de vista académico, el trabajo refuerza la idea de que la vulnerabilidad financiera no debe analizarse solo desde la renta, sino desde el equilibrio presupuestario y la capacidad degeneración de excedente. Contribuye a la literatura sobre medición multidimensional de la pobreza y sobre salud financiera, y muestra la utilidad de técnicas de aprendizaje automático (clustering y árboles de decisión) en el análisis de microdatos de hogares. Desde el punto de vista empresarial, la identificación de perfiles y la posibilidad de predecirlos con modelos interpretables resulta estratégica para una entidad como MAPFRE en gestión del riesgo, diseño de productos adaptados a distintos segmentos y segmentación avanzada de clientes. En un entorno regulatorio cada vez más exigente en materia de conducta de negocio y protección al consumidor, disponer de indicadores de salud financiera basados en datos y en metodología transparente puede mejorar la adecuación de los productos y la gestión del riesgo de impago.

Un aspecto relevante de los resultados es que, pese a trabajar con una población seleccionada por presentar mala salud financiera en el proyecto global, la dimensión específica de ahorro/ingresos identifica un peso considerable del perfil etiquetado como de "alta capacidad financiera". Esta aparente paradoja se explica porque la clasificación se realiza exclusivamente en función del ratio ahorro/ingresos y es, por tanto, relativa al conjunto de hogares analizados. Es decir, el perfil de alta capacidad financiera no agrupa necesariamente a hogares sin problemas económicos en términos absolutos, sino a aquellos que, dentro de una población ya vulnerable por otras dimensiones, muestran un excedente presupuestario mayor que el resto. En muchos casos, se trata de hogares con ingresos moderados que han ajustado su estructura de gasto y consiguen mantener un superávit, mientras que otros hogares con características socioeconómicas similares permanecen en déficit estructural. Este resultado refuerza la idea de que la salud financiera es un fenómeno multidimensional: un hogar puede situarse en el perfil "alto" de la dimensión ahorro/ingresos

y, sin embargo, seguir presentando fragilidades importantes en otras dimensiones (esfuerzo en vivienda, carga de deuda, privación material), lo que justifica la necesidad de integrar todas las dimensiones en un score global.

Como limitaciones y líneas futuras cabe citar: (i) la dependencia de la EPF en cuanto a definiciones, periodicidad y cobertura; (ii) la necesidad de integrar esta dimensión con el resto del score de salud financiera (vivienda, energía, privación material) para obtener una visión global; (iii) la posible extensión a otros modelos supervisados (redes neuronales, modelos ensemble) y la comparación sistemática de su rendimiento y interpretabilidad; (iv) el análisis de sensibilidad temporal, replicando la segmentación en varios años para evaluar la estabilidad de los perfiles; (v) la incorporación, cuando esté disponible, de información sobre stock de deuda y activos para enriquecer el indicador de capacidad de ahorro. En conjunto, la dimensión de capacidad de ahorro se configura como un componente esencial del modelo global de salud financiera, con rigor metodológico y aplicabilidad práctica tanto en el ámbito académico como en el empresarial.

Capítulo 7. BIBLIOGRAFÍA

[1] EuroStat, «People at risk of poverty or social exclusion in 2023». [En línea]. Disponible en: https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20240612-1

[2] Banco de España, «Informe de la Situación Financiera de los hogares y las empresas». [En línea]. Disponible en: https://www.bde.es/f/webbe/SES/Secciones/Publicaciones/Informesituacionfinancierafamiliasyempresas/2023/S2/Fich/SituacionFinanciera_022023.pdf

[3] Sabina Alkire y James Foster, «Counting and multidimensional poverty measurement». [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S0047272710001660

[4] A. B. Atkinson, «On the Measurement of Poverty». [En línea]. Disponible en: https://www.jstor.org/stable/1911028

[5] Consumer Financial Protection Bureau, «Financial well-being: The goal of financial education». [En línea]. Disponible en: https://files.consumerfinance.gov/f/201501_cfpb_report_financial-well-being.pdf

[6] Olivia S. Mitchell y Annamaria Lusardi, «Financial Literacy and Economic Outcomes: Evidence and Policy Implications». [En línea]. Disponible en: https://gflec.org/wp-content/uploads/2015/12/Financial-Literacy-and-Economic-Outcomes-Journal-of-Retirement-Summer-2015.pdf

[7] Hastie Trevor, Robert Tibshirani, y Jerome Friedman, «The Elements of Statistical Learning: Data Mining, Inference and Prediction». [En línea]. Disponible en: https://hastie.su.domains/ElemStatLearn/

[8] Hal R. Varian, «Big Data: New Tricks for Econometrics». [En línea]. Disponible en: https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3

[9] Susan Athey y Guido W. Imbens, «The State of Applied Econometrics: Causality and Policy Evaluation». [En línea]. Disponible en: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.3

[10] Brian Nolan y Christopher T. Whelan, «Using non-monetary deprivation indicators to analyze poverty and social exclusion: Lessons from Europe?» [En línea]. Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/pam.20493

[11] «El algoritmo k-means aplicado a clasificación y procesamiento de imágenes».

[12] A. K. Jain, «Data clustering: 50 years beyond K-means». [En línea]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0167865509002323

ANEXO I

FRAGMENTO DE CÓDIGO RELEVANTE (PYTHON)

En este fragmento se muestra la parte esencial del código utilizada para construir el ratio ahorro/ingresos, realizar la segmentación K-means y entrenar el árbol de decisión sobre la base de datos de la EPF:

```
#Selección de variables

X = datos_final[[

   "Tamaño del hogar",

   "densidad de población",

   "CC. AA."

]].copy()

y = datos_final["perfil_ahorro"]

# Codificar variables categóricas

X = pd.get_dummies(X, drop_first=True)



# Train / Test Split

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(

   X,
```

```
    y,
    test_size=0.3,
    random_state=42,
    stratify=y
)

# Entrenar Árbol de Decisión
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, accuracy_score
model_tree = DecisionTreeClassifier(
    max_depth=4,
    random_state=42
)
model_tree.fit(X_train, y_train)

# Evaluación
y_pred = model_tree.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n")
print(classification_report(y_test, y_pred))
```

```
# Importancia de variables

import pandas as pd

importances = pd.Series(

   model_tree.feature_importances_,

   index=X.columns

).sort_values(ascending=False)

print("\nImportancia de variables:\n")

print(importances)
```

```
Accuracy: 0.6842105263157895

Classification Report:

                          precision    recall  f1-score   support

       Ahorrador moderado       0.71      0.84      0.77       147
Alta capacidad financiera       0.67      0.20      0.31        59
      Déficit estructural       0.63      0.78      0.70        60

                 accuracy                           0.68       266
                macro avg       0.67      0.61      0.59       266
             weighted avg       0.68      0.68      0.65       266


Importancia de variables:

CC. AA.                 0.571006
Tamaño del hogar        0.253117
densidad de población   0.175877
dtype: float64
```

Figura 14 - Resultados arbol de decisión

DOCUMENTACIÓN Y REFERENCIAS ESTADÍSTICAS

La Encuesta de Presupuestos Familiares (EPF) es una operación estadística del Instituto Nacional de Estadística (INE) que se realiza de forma continua desde 2006, con

metodología armonizada a nivel europeo (Reglamento CE 2454/97 y sucesivas modificaciones). La documentación metodológica, los cuestionarios y las notas de prensa están disponibles en la web del INE. Los microdatos se obtienen mediante solicitud al INE o a través del Banco de Datos del INE cuando están en formato de uso público. Para la construcción de estimaciones representativas de la población, deben utilizarse los factores de elevación (ponderadores) que proporciona la encuesta; en el presente trabajo, el análisis se ha realizado sobre la muestra de hogares con información completa, y la generalización de los porcentajes de cada perfil a la población requeriría la aplicación de los pesos muestrales correspondientes. Se recomienda consultar las publicaciones del INE sobre la EPF y las de Eurostat sobre las encuestas de presupuestos familiares armonizadas (Household Budget Survey, HBS) para detalles sobre definiciones, cobertura y limitaciones de la encuesta.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, María Sarrado Vázquez, estudiante del Doble Grado de Ingeniería de Telecomunicaciones y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado “La salud financiera de los hogares españoles: evidencia a partir de la Encuesta de Presupuestos Familiares", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.
2. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
5. Interpretador de código: Para realizar análisis de datos preliminares.
6. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
7. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.

8. Generador de datos sintéticos de prueba: Para la creación de conjuntos de datos ficticios.
9. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 20/04/2026

Firma: 