



Facultad de Ciencias Económicas y Empresariales

**PERSISTENCIA Y TRANSICIONES DEL RIESGO DE POBREZA O
EXCLUSIÓN SOCIAL EN ESPAÑA: EVIDENCIA CON DATOS DE PANEL
ECV (EU-SILC).**

María Caballero Pascual

5º E-3 Analytics

Director: Manuel Alejandro Betancourt Odio

ÍNDICE

RESUMEN.....	5
ABSTRACT	6
INTRODUCCIÓN	7
CAPÍTULO I: EL INDICADOR AROPE Y LA MEDICIÓN DE LA POBREZA EN EUROPA	11
1.1. De la pobreza monetaria a la exclusión social: evolución del marco conceptual	11
1.2. El indicador AROPE: definición, origen y marco institucional.....	12
1.3. Componentes del indicador AROPE	12
1.3.1. <i>Riesgo de pobreza monetaria</i>	13
1.3.2. <i>Privación material y social severa</i>	13
1.3.3. <i>Baja intensidad laboral del hogar</i>	14
1.4. Ventajas, limitaciones y debates en torno al indicador AROPE.....	15
1.5. Justificación del uso del AROPE y enfoque longitudinal de este trabajo.....	16
CAPÍTULO II: DATOS, FUENTES Y METODOLOGÍA	17
2.1. Fuente de datos: la Encuesta de Condiciones de Vida (ECV/EU-SILC)	17
2.2. Integración de las bases de datos y construcción del dataset	18
2.3. Depuración y tratamiento de los datos	19
2.4. Construcción del indicador AROPE.....	20
2.4.1. <i>Componente de pobreza monetaria</i>	20
2.4.2. <i>Componente de privación material y social severa</i>	21
2.4.3. <i>Componente de baja intensidad laboral</i>	21
2.5. Construcción del indicador final y validación.....	21
2.6. Muestra final y estructura del panel	22
2.7. Variables utilizadas en el análisis.....	23
2.8. Métodos de análisis	25
CAPÍTULO III: ANÁLISIS DESCRIPTIVO Y LONGITUDINAL DEL AROPE EN ESPAÑA (2020–2023)	26
3.1. Construcción de la muestra de análisis.....	26
3.2. Análisis descriptivo estático: distribución del AROPE por características sociodemográficas.....	27
3.3. Análisis longitudinal: persistencia y transiciones del AROPE (2020–2023).....	30
3.3.1. <i>Evolución del AROPE en el panel 2020–2023</i>	30
3.3.2. <i>Patrones de persistencia</i>	31
3.3.3. <i>Matrices de transición año a año</i>	32
3.3.4. Perfil sociodemográfico según tipo de trayectoria	33
3.4. Síntesis del análisis descriptivo y longitudinal.....	34

CAPÍTULO IV: MODELIZACIÓN PREDICTIVA MEDIANTE APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)	36
4.1. Justificación de la aproximación algorítmica	36
4.2. Fundamentos del algoritmo Random Forest	37
4.3. Especificación técnica y entrenamiento del modelo	38
4.4. Evaluación del rendimiento y capacidad de detección.....	39
4.5. Jerarquía de determinantes del riesgo AROPE	40
4.6. Síntesis del capítulo	43
CAPÍTULO V: CONCLUSIONES.....	44
5.1 Sobre la distribución y los determinantes estáticos del riesgo AROPE.....	44
5.2 Sobre la persistencia y las transiciones del riesgo AROPE	45
5.3 Sobre el modelo predictivo y su potencial para la detección temprana	46
5.4 Limitaciones del trabajo y líneas de investigación futura	47
REFERENCIAS.....	48

RESUMEN

El presente Trabajo Fin de Grado analiza el fenómeno del riesgo de pobreza o exclusión social en España, medido a través del indicador AROPE (*At Risk of Poverty or Social Exclusion*), utilizado por la Unión Europea para evaluar la vulnerabilidad social desde una perspectiva multidimensional. El objetivo principal es identificar qué factores económicos, laborales, educativos y sociodemográficos se asocian en mayor medida a este riesgo, así como explorar si es posible anticipar situaciones de vulnerabilidad mediante técnicas de análisis predictivo.

Para ello, se utilizan los microdatos de la Encuesta de Condiciones de Vida, integrada en el programa europeo EU-SILC. A partir de estos datos, se reconstruye el indicador AROPE, combinando sus tres dimensiones: riesgo de pobreza monetaria, privación material y social severa, y baja intensidad laboral del hogar. Posteriormente, se realiza un análisis descriptivo de las principales variables asociadas al riesgo de pobreza y se aplica un modelo de aprendizaje automático, concretamente un Random Forest, con el fin de identificar patrones predictivos y determinar la importancia relativa de cada variable.

Los resultados muestran que el riesgo de pobreza o exclusión social no se distribuye de forma homogénea entre la población española. Los perfiles más vulnerables se concentran entre personas con menor nivel educativo, jóvenes, hogares numerosos y personas vinculadas a situaciones laborales más inestables o precarias. Asimismo, el modelo predictivo identifica como variables relevantes la privación material severa, la edad, la educación, la situación laboral y el tipo de contrato.

En conjunto, el trabajo muestra que la pobreza y la exclusión social son fenómenos complejos, condicionados por la interacción de factores económicos, laborales y sociales, y que las técnicas de machine learning pueden complementar el análisis tradicional para mejorar la detección temprana de situaciones de vulnerabilidad.

Palabras clave: Pobreza; exclusión social; AROPE; EU-SILC; vulnerabilidad social; machine learning.

ABSTRACT

This Final Degree Project analyses the phenomenon of the risk of poverty or social exclusion in Spain, measured through the AROPE indicator (*At Risk of Poverty or Social Exclusion*), used by the European Union to assess social vulnerability from a multidimensional perspective. The main objective is to identify the economic, labour, educational and sociodemographic factors most strongly associated with this risk, as well as to explore whether vulnerable situations can be anticipated through predictive analysis techniques.

To this end, the project uses microdata from the Living Conditions Survey, integrated into the European EU-SILC programme. Based on these data, the AROPE indicator is reconstructed by combining its three dimensions: risk of monetary poverty, severe material and social deprivation, and low household work intensity. A descriptive analysis of the main variables associated with the risk of poverty is then carried out, followed by the application of a machine learning model, specifically a Random Forest, in order to identify predictive patterns and determine the relative importance of each variable.

The results show that the risk of poverty or social exclusion is not evenly distributed among the Spanish population. The most vulnerable profiles are concentrated among people with lower educational attainment, young people, large households, and individuals linked to more unstable or precarious employment situations. Likewise, the predictive model identifies severe material deprivation, age, education, labour status and type of contract as relevant variables.

Overall, this project shows that poverty and social exclusion are complex phenomena shaped by the interaction of economic, labour and social factors, and that machine learning techniques can complement traditional analysis to improve the early detection of vulnerable situations.

Keywords: Poverty; social exclusion; AROPE; EU-SILC; social vulnerability; machine learning.

INTRODUCCIÓN

La pobreza y la exclusión social constituyen uno de los principales desafíos estructurales del proyecto europeo contemporáneo. En el marco de la Unión Europea, la medición oficial de este fenómeno se articula a través del indicador AROPE (*At Risk of Poverty or Social Exclusion*), una medida compuesta que integra tres dimensiones complementarias: el riesgo de pobreza monetaria, la privación material y social severa, y la baja intensidad laboral del hogar. Este indicador no agrega mecánicamente sus componentes, sino que considera a cada persona una única vez aunque cumpla simultáneamente varias condiciones, lo que lo convierte en una medida de la unión —y no de la suma— de las situaciones de vulnerabilidad (Eurostat, 2025).

El AROPE se ha consolidado como el principal instrumento estadístico para el seguimiento de los compromisos europeos en materia de reducción de la pobreza, desde la Estrategia Europa 2020 hasta los objetivos del Pilar Europeo de Derechos Sociales y la Agenda 2030. En España, los datos disponibles reflejan una incidencia elevada y persistente: según el seguimiento del indicador realizado por la EAPN-España, en 2018 se estimaba que 12.188.288 personas —el 26,1% de la población— se encontraban en situación AROPE (Llano Ortiz, 2019). Los informes de seguimiento más recientes del Ministerio de Derechos Sociales confirman que, si bien el indicador ha registrado mejoras en el período 2019–2024, estas han sido desiguales según el componente analizado y la comunidad autónoma considerada (Ministerio de Derechos Sociales, 2024).

Qué sabemos. La investigación académica y los informes institucionales han identificado con consistencia los principales factores asociados al riesgo AROPE en España y en el conjunto de la Unión Europea. El nivel educativo, la situación laboral, la edad y la estructura del hogar aparecen sistemáticamente como determinantes relevantes de la vulnerabilidad. Desde una perspectiva macroeconómica, el Banco de España ha documentado cómo la desigualdad en la distribución de la renta, el consumo y la riqueza configura un escenario de oportunidades muy diferenciadas entre grupos sociales (Banco de España, 2018). A nivel comparativo europeo, estudios recientes han aplicado metodologías de lógica difusa al AROPE con el objetivo de capturar la intensidad relativa de la desventaja multidimensional, superando la visión binaria del indicador oficial (García-Pardo, Bárcena-Martín y Pérez-Moreno, 2020). En el plano territorial, la investigación sobre la evolución regional del AROPE en España concluye que la crisis

económica no modificó sustancialmente el patrón territorial del riesgo de pobreza, aunque sí afectó de forma diferenciada a las comunidades autónomas, siendo la tasa de pobreza monetaria el componente con mayor capacidad explicativa de la variabilidad regional (Lafuente Lechuga, García Luque y Faura Martínez, 2015).

El análisis comparado entre países miembros de la Unión Europea muestra que el riesgo de pobreza y exclusión responde a estructuras institucionales, mercados laborales y sistemas de protección social con características propias, lo que hace de España un caso de especial interés por la combinación de una elevada tasa de temporalidad laboral, un sistema de transferencias sociales de cobertura media y una fuerte heterogeneidad regional en las condiciones de vida (Podadera Rivera, Calderón Milán y García Perea, 2020). Los informes territoriales más recientes confirman que esta heterogeneidad persiste: incluso en comunidades con niveles de renta media relativamente elevados pueden mantenerse tasas significativas de exclusión asociadas a desigualdad interna y a una limitada capacidad redistributiva de las transferencias públicas (EAPN-ES, 2023).

Qué no sabemos. Sin embargo, tanto los informes institucionales de seguimiento como los análisis académicos revisados adolecen de una limitación estructural común: su enfoque es predominantemente transversal. La mayoría de los estudios sobre el AROPE en España analizan tasas anuales, comparaciones territoriales o perfiles sociodemográficos en un momento del tiempo, sin explotar de forma sistemática la dimensión longitudinal de la Encuesta de Condiciones de Vida. Esta laguna analítica es relevante porque la pobreza puede presentar dinámicas muy distintas: hay episodios breves asociados a perturbaciones coyunturales —la pérdida temporal del empleo, un gasto extraordinario imprevisto— y hay trayectorias prolongadas ancladas en factores estructurales como la baja cualificación, la precariedad contractual crónica o el aislamiento social. Confundir ambas no solo empobrece el diagnóstico, sino que puede conducir a intervenciones de política pública inadecuadas: una medida de activación laboral es eficaz para quien atraviesa un episodio puntual de riesgo, pero insuficiente para quien lleva años en una situación de exclusión estructural.

La Encuesta de Condiciones de Vida tiene una estructura de panel rotatorio que permite seguir a los mismos individuos durante cuatro años consecutivos. Esta característica metodológica —que Eurostat señala como fundamental para el cálculo del indicador de pobreza persistente (Eurostat, 2025)— ha sido sistemáticamente infrautilizada en los análisis del AROPE para España. El INE publica con regularidad estadísticas descriptivas

sobre la evolución anual del indicador (INE, 2025), pero el análisis de las trayectorias individuales, de las probabilidades de transición entre estados de riesgo y de los factores que determinan la duración de los episodios de pobreza sigue siendo una asignatura pendiente en la literatura académica española.

Qué aporta este trabajo. El presente Trabajo Fin de Grado aborda directamente esa laguna. A partir de los microdatos de la ECV para los años 2020–2023, se reconstruye el indicador AROPE y se explota la estructura de panel rotatorio de la encuesta para analizar no solo quién está en riesgo en un momento dado, sino durante cuánto tiempo y con qué trayectoria. Los datos revelan que, entre los individuos observados en los cuatro años del panel, el 9,9% permanece en situación AROPE de forma ininterrumpida —lo que cabe calificar como pobreza persistente en sentido estricto—, mientras que una proporción significativa experimenta transiciones de entrada o salida del riesgo que responden a factores identificables. Estos patrones, invisibles para el análisis transversal convencional, constituyen el núcleo empírico de este trabajo.

Complementariamente, el trabajo incorpora un modelo de aprendizaje automático basado en el algoritmo *Random Forest*, que frente a los métodos estadísticos tradicionales permite detectar relaciones no lineales e interacciones entre variables y jerarquizar los determinantes del riesgo con mayor capacidad predictiva. La combinación del análisis longitudinal con la modelización predictiva ofrece una visión más completa del fenómeno: el primero describe quién está en riesgo y por cuánto tiempo; el segundo permite anticipar qué características individuales predicen con mayor fuerza tanto la entrada como la permanencia en situación de exclusión.

El objetivo principal del trabajo es, por tanto, doble: caracterizar la persistencia y las transiciones del AROPE en España a partir de microdatos de panel individuales, e identificar los factores sociodemográficos y laborales que con mayor fuerza predicen tanto la permanencia en el riesgo como la salida del mismo. Con ello, el trabajo aspira a contribuir al debate académico sobre la dinámica de la exclusión social en España y a ofrecer una base empírica útil para el diseño de políticas de detección temprana y prevención de la pobreza persistente.

El resto del trabajo se estructura del siguiente modo. El Capítulo I presenta el marco conceptual del indicador AROPE, sus componentes, ventajas, limitaciones y el debate académico en torno a su uso. El Capítulo II describe la fuente de datos y el proceso de

construcción del indicador y de la base analítica. El Capítulo III desarrolla el análisis descriptivo y longitudinal: distribución del riesgo por características sociodemográficas, patrones de persistencia y matrices de transición. El Capítulo IV presenta el modelo predictivo *Random Forest* y sus resultados. El Capítulo V recoge las conclusiones y el Capítulo VI las recomendaciones de política pública derivadas de los hallazgos empíricos.

CAPÍTULO I: EL INDICADOR AROPE Y LA MEDICIÓN DE LA POBREZA EN EUROPA

1.1. De la pobreza monetaria a la exclusión social: evolución del marco conceptual

La pobreza es un fenómeno complejo que no puede ser entendido únicamente como la falta de ingresos. Tradicionalmente, el análisis económico ha utilizado medidas monetarias para identificar a las personas en situación de pobreza. Sin embargo, este enfoque resulta limitado, ya que no tiene en cuenta otros aspectos relevantes como las condiciones de vida, el acceso a bienes básicos o la participación en el mercado laboral.

La crítica más influyente a esta reducción de la pobreza al ingreso proviene del trabajo de Amartya Sen, quien argumentó que lo relevante no es lo que los individuos poseen, sino lo que son capaces de hacer y ser con esos recursos (Sen, 1983). Desde este enfoque de las capacidades (*capability approach*), una persona puede disponer de ingresos suficientes y, sin embargo, verse privada de capacidades fundamentales por razones de salud, discriminación, exclusión del mercado laboral o aislamiento social. Esta perspectiva desplazó el debate hacia una concepción multidimensional de la pobreza, en la que los ingresos constituyen solo una de las dimensiones relevantes.

En este contexto, el concepto de exclusión social surge como una ampliación del enfoque tradicional, incorporando dimensiones adicionales que permiten capturar de forma más completa la situación de vulnerabilidad de los individuos. La exclusión social alude a la incapacidad de participar plenamente en la vida económica, social y cívica de la comunidad, e incorpora factores como el empleo, la educación, la vivienda o las redes de apoyo (Room, 1995). De este modo, la pobreza pasa a entenderse como un fenómeno multidimensional, en el que intervienen factores económicos, sociales y laborales. Este enfoque ha sido adoptado por la Unión Europea en el marco de sus políticas sociales, dando lugar al desarrollo de indicadores que permiten medir de forma más precisa el riesgo de pobreza y exclusión social.

1.2. El indicador AROPE: definición, origen y marco institucional

El indicador AROPE (*At Risk of Poverty or Social Exclusion*) es una medida desarrollada por la Unión Europea para evaluar el nivel de pobreza y exclusión social en los Estados miembros. Este indicador fue introducido en el marco de la Estrategia Europa 2020, cuyo objetivo era reducir el número de personas en riesgo de pobreza o exclusión social en al menos 20 millones. Actualmente, el AROPE sigue siendo el indicador de referencia para monitorizar los objetivos sociales de la Agenda Europa 2030 y el Pilar Europeo de Derechos Sociales (Eurostat, 2025).

Según la definición de Eurostat (2025), el indicador AROPE mide el porcentaje de población que se encuentra en al menos una de las siguientes situaciones: riesgo de pobreza monetaria, privación material y social severa, o baja intensidad laboral en el hogar. Una característica metodológica esencial es que cada persona se contabiliza una única vez, aunque cumpla simultáneamente más de una de estas condiciones. Ello implica que el AROPE es una medida de la unión de las tres dimensiones, no de su suma, lo que garantiza que no se produzca doble contabilización, pero también que se agreguen bajo una misma categoría binaria situaciones de vulnerabilidad muy heterogéneas.

Este indicador se calcula a partir de los datos recogidos en la encuesta EU-SILC (*European Union Statistics on Income and Living Conditions*), que constituye la principal fuente estadística para el análisis de las condiciones de vida en Europa. Las bases conceptuales del sistema de indicadores sociales europeo del que forma parte el AROPE fueron establecidas en los trabajos de Atkinson, Cantillon, Marlier y Nolan (2002; 2004), que sentaron los principios de validez, fiabilidad, comparabilidad e interpretabilidad para el diseño de indicadores de inclusión social en la UE.

1.3. Componentes del indicador AROPE

El indicador AROPE se construye a partir de tres dimensiones que reflejan distintos aspectos de la pobreza y la exclusión social. La inclusión de estas tres dimensiones

permite captar la naturaleza multidimensional del fenómeno, superando las limitaciones de los indicadores puramente monetarios.

1.3.1. Riesgo de pobreza monetaria

El primer componente del indicador se refiere al riesgo de pobreza basado en ingresos. Según Eurostat (2025), una persona se considera en riesgo de pobreza cuando su renta disponible equivalente se sitúa por debajo del 60% de la mediana nacional. Este indicador es de carácter relativo, lo que significa que mide la posición de los individuos en relación con el conjunto de la sociedad, y no refleja necesariamente la falta absoluta de recursos, sino la desigualdad en la distribución de la renta.

Esta naturaleza relativa tiene implicaciones metodológicas importantes: el umbral se ajusta automáticamente con el crecimiento económico, de modo que una mejora generalizada de las rentas no reduce necesariamente la tasa de pobreza si la distribución de la renta no se modifica. El propio Eurostat (2025) reconoce que el riesgo de pobreza monetaria en un momento dado no implica necesariamente bajos niveles de vida a corto plazo, ya que los individuos pueden recurrir a ahorros, crédito o redes de apoyo familiar para amortiguar temporalmente su situación. Esta observación es especialmente pertinente para el análisis de la dinámica del indicador: la capacidad de un individuo para sostenerse durante un episodio breve de pobreza monetaria puede ser muy diferente de su situación ante una trayectoria prolongada de riesgo, aspecto central que este trabajo aborda.

1.3.2. Privación material y social severa

El segundo componente del indicador AROPE mide la privación material, es decir, la incapacidad de los hogares para acceder a determinados bienes o servicios considerados básicos. Según la metodología de Eurostat (2025), actualizada en 2021 para incorporar la dimensión social a la estrictamente material, una persona se encuentra en situación de privación material y social severa cuando no puede permitirse al menos cinco de los trece ítems de una lista armonizada, que incluye aspectos como mantener la vivienda a una

temperatura adecuada, hacer frente a gastos imprevistos, disponer de determinados bienes duraderos, o participar en actividades sociales y de ocio.

Este indicador permite captar situaciones de vulnerabilidad que no siempre se reflejan en los ingresos, incorporando una dimensión más directa de las condiciones de vida. Además, resulta especialmente sensible a la pobreza persistente: un hogar puede superar el umbral de renta y, sin embargo, arrastrar carencias materiales acumuladas durante periodos anteriores de dificultad. Por ello, su análisis longitudinal, tal y como se aborda en este trabajo, resulta especialmente revelador.

1.3.3. Baja intensidad laboral del hogar

El tercer componente del indicador AROPE mide la intensidad laboral del hogar. Según Eurostat (2025), este indicador se define como la proporción de meses trabajados por los miembros del hogar en edad laboral (18-59 años, excluidos los estudiantes de 18-24 años) en relación con el total de meses potenciales. Un hogar presenta baja intensidad laboral cuando este cociente es inferior al 20%.

Este componente resulta especialmente relevante, ya que la falta de empleo o la baja intensidad de trabajo constituyen factores clave en la generación de situaciones de pobreza. Además, conecta la situación de pobreza con la estructura de empleo del hogar, incorporando una dimensión dinámica: el empleo no solo determina el nivel de ingresos presentes, sino también la acumulación de derechos sociales —pensiones, prestaciones por desempleo— que condicionan la vulnerabilidad futura. La investigación sobre persistencia de la pobreza en España ha puesto de manifiesto que el empleo, la cualificación educativa y la composición del hogar son los factores que con mayor fuerza predicen tanto la entrada como la permanencia en situación de riesgo (Ayllón, 2013).

1.4. Ventajas, limitaciones y debates en torno al indicador AROPE

El indicador AROPE presenta varias ventajas frente a las medidas tradicionales de pobreza basadas únicamente en ingresos. En primer lugar, permite incorporar distintas dimensiones del bienestar, ofreciendo una visión más completa de la situación de los individuos. En segundo lugar, facilita la comparación entre países, ya que se basa en una metodología armonizada a nivel europeo (Atkinson et al., 2002). Por último, resulta especialmente útil para el diseño de políticas públicas, al identificar diferentes tipos de vulnerabilidad que requieren intervenciones distintas.

Sin embargo, el indicador también presenta limitaciones de fondo que la literatura ha señalado sistemáticamente. En primer lugar, su umbral monetario relativo puede resultar paradójico: en períodos de contracción económica generalizada, la mediana de ingresos cae y con ella el umbral de pobreza, de modo que algunas personas "salen" estadísticamente de la pobreza no porque su situación mejore, sino porque el conjunto de la sociedad empobrece. El propio Eurostat (2025) reconoce la arbitrariedad en la elección de los umbrales y calcula indicadores complementarios con umbrales alternativos para ofrecer una visión más robusta.

En segundo lugar, el AROPE es un indicador esencialmente estático: mide el estado de cada individuo en el año de referencia, pero no informa sobre cuánto tiempo lleva en esa situación ni sobre las trayectorias que lo han llevado hasta ella. Esta limitación es señalada explícitamente por Eurostat (2025), que advierte que "el impacto acumulado de períodos prolongados en riesgo debe evaluarse con más detalle". Esta es, precisamente, la laguna analítica que motiva el enfoque longitudinal de este trabajo.

En tercer lugar, como medida de unión de tres componentes heterogéneos, el AROPE agrega bajo una misma categoría situaciones de vulnerabilidad cualitativamente distintas. Frente a estas críticas, enfoques alternativos como los índices de pobreza multidimensional de tipo Alkire-Foster o las aproximaciones de conjuntos difusos (*fuzzy sets*) intentan capturar la intensidad relativa de la desventaja (Atkinson et al., 2002), pero a costa de mayor complejidad metodológica y menor comparabilidad institucional. El

AROPE, con sus limitaciones, sigue siendo el estándar de referencia para el seguimiento de la pobreza en Europa.

1.5. Justificación del uso del AROPE y enfoque longitudinal de este trabajo

El uso del indicador AROPE en este trabajo se justifica por su capacidad para capturar de forma integral el riesgo de pobreza y exclusión social. Dado que el objetivo del estudio es analizar la persistencia y las transiciones en este riesgo, resulta necesario utilizar una medida que incorpore múltiples dimensiones. Asimismo, el hecho de que el indicador esté basado en la encuesta EU-SILC permite trabajar con datos comparables y ampliamente utilizados en la literatura académica y en el ámbito institucional.

La EU-SILC tiene una estructura de panel rotatorio en la que cada submuestra sigue a los mismos individuos durante cuatro años consecutivos (Eurostat, 2025). El componente longitudinal de la encuesta se centra precisamente en el estudio de los cambios individuales a lo largo del tiempo —transiciones en los niveles de renta y pobreza, en la situación laboral y educativa— y desempeña un papel crucial en el cálculo del indicador de pobreza persistente. Esta característica, frecuentemente infrautilizada en los análisis del AROPE para España, es la que permite abordar la pregunta central de este trabajo: no solo *quién* está en riesgo, sino *durante cuánto tiempo y con qué trayectoria*.

En este trabajo, el indicador AROPE se ha reconstruido a partir de los microdatos de la ECV para los años 2020–2023, lo que permite un control completo sobre el proceso de cálculo y facilita su utilización en el análisis empírico posterior. Adicionalmente, se explota la estructura de panel para analizar patrones de persistencia y transición: los datos revelan que, entre los individuos observados durante los cuatro años del periodo analizado, el 9,9% se encuentra en situación AROPE de forma continuada —lo que puede calificarse como pobreza persistente en sentido estricto—, mientras que una proporción significativa experimenta entradas y salidas del riesgo que responden a factores identificables. Este enfoque permite superar la principal limitación del AROPE como indicador estático y ofrece una visión más completa y accionable de la dinámica de la exclusión social en España.

CAPÍTULO II: DATOS, FUENTES Y METODOLOGÍA

Este capítulo describe la fuente de datos, el proceso de construcción de la base analítica y las decisiones metodológicas que sustentan el análisis empírico del trabajo. Se estructura en seis apartados: la descripción de la ECV/EU-SILC y su estructura longitudinal, el proceso de integración y depuración de ficheros, la construcción del indicador AROPE, la caracterización de la muestra final, la descripción de las variables utilizadas y la justificación de los métodos de análisis aplicados.

2.1. Fuente de datos: la Encuesta de Condiciones de Vida (ECV/EU-SILC)

El análisis desarrollado en este trabajo se basa en los microdatos de la Encuesta de Condiciones de Vida (ECV), la versión española de la encuesta europea EU-SILC (*European Union Statistics on Income and Living Conditions*). Esta encuesta, coordinada por Eurostat y ejecutada en España por el Instituto Nacional de Estadística (INE), constituye la principal fuente estadística armonizada para el estudio de la pobreza y la exclusión social en la Unión Europea (Eurostat, 2025).

La ECV recoge información anual sobre los ingresos, las condiciones de vida y la situación laboral de individuos y hogares. Una de sus características más relevantes para este trabajo es su estructura de panel rotatorio: cada submuestra sigue a los mismos individuos durante cuatro años consecutivos, de modo que en cada oleada anual conviven cuatro cohortes en distintas etapas del seguimiento. Esta estructura permite explotar tanto la dimensión transversal —analizando la situación en un año concreto— como la dimensión longitudinal —rastreamiento de la evolución de los individuos a lo largo del tiempo—, siendo esta última la que posibilita el análisis de persistencia y transiciones del AROPE que constituye el núcleo de este trabajo (Eurostat, 2025).

El período analizado comprende los años 2020 a 2023, cubriendo tanto el impacto inmediato de la pandemia de COVID-19 sobre las condiciones de vida de los hogares españoles como la fase de recuperación económica posterior. Este período ofrece una ventana de observación especialmente rica para el estudio de la dinámica del riesgo AROPE, al incluir episodios tanto de deterioro como de mejora de las condiciones socioeconómicas.

2.2. Integración de las bases de datos y construcción del dataset

La información de la ECV se distribuye en cuatro ficheros diferenciados según el nivel de análisis. La Tabla 1 resume su contenido y las variables clave utilizadas en este trabajo.

Tabla 1. Ficheros de la ECV y variables utilizadas en el análisis.

Fichero ECV	Contenido principal	Variables clave utilizadas
Fichero de personas (P)	Características individuales: edad, sexo, estado civil, nivel educativo, situación laboral	EDAD, PB150, PB190, PE040, PE041, PL031, PL040, PL140, PL145
Fichero de hogares (H)	Características del hogar: tamaño, composición, ítems de privación material	TAMANO_HOGAR_GRUPO, MENORES, ítems de privación (HD010–HD080)
Fichero de renta (R)	Ingresos individuales y del hogar por fuentes (empleo, prestaciones, pensiones, otros)	PY010N–PY140N, HY010–HY145

Fichero longitudinal (L)	Seguimiento de individuos a lo largo de los años del panel rotatorio	PB030 (identificador personal), PB010 (año de referencia)
---------------------------------	--	--

Nota: Elaboración propia a partir de la documentación metodológica de la ECV (INE, 2020–2023) y de Eurostat (2025).

El proceso de construcción del dataset ha requerido la integración de estos ficheros a partir de identificadores únicos. Para el fichero de personas se ha utilizado la clave KEY_PER_ANO, que combina el identificador personal (PB030) con el año de referencia (PB010), garantizando la unicidad de cada observación individuo-año. Para el fichero de hogares se ha empleado la clave KEY_HOGAR_ANO, que combina el identificador del hogar con el año. La vinculación entre ficheros se ha realizado mediante funciones de búsqueda (BUSCARX, ÍNDICE-COINCIDIR) en Excel, trasladando las variables del hogar y de renta al fichero de personas como unidad de análisis principal.

El resultado es un dataset longitudinal en el que cada observación representa a un individuo en un año determinado, incluyendo sus características personales, las del hogar al que pertenece y su situación de ingresos. La variable longitudinal clave es el identificador personal PB030, que permanece constante para cada individuo a lo largo de los cuatro años del panel y permite rastrear su trayectoria respecto al indicador AROPE.

2.3. Depuración y tratamiento de los datos

Tras la integración de los ficheros, se ha llevado a cabo un proceso sistemático de depuración. La ECV utiliza códigos específicos para indicar valores no disponibles, no aplicables o respuestas incompletas, que deben identificarse y tratarse adecuadamente antes del análisis.

En concreto, se han sustituido por valores perdidos (missing values) los códigos que la ECV asigna a las categorías de respuesta "no consta" o "no aplicable" en las variables de situación laboral (PL031, PL040, PL140, PL145) y nivel educativo (PE040, PE041). Las variables de renta con valores negativos o extremadamente elevados —que en algunos

casos reflejan errores de codificación o situaciones atípicas— han sido tratadas con especial cuidado en la construcción del componente de pobreza monetaria. Las observaciones con valores perdidos en alguna de las variables del modelo predictivo han sido eliminadas de la submuestra utilizada para el Random Forest (N final = 7.590 individuos), manteniéndose en cambio para el análisis descriptivo y longitudinal.

2.4. Construcción del indicador AROPE

El indicador AROPE se ha reconstruido desde los microdatos siguiendo la metodología oficial de Eurostat (2025). La construcción se ha realizado en tres etapas correspondientes a sus tres componentes, que se detallan a continuación.

2.4.1. Componente de pobreza monetaria

El primer componente identifica a los individuos cuya renta disponible equivalente por unidad de consumo se sitúa por debajo del 60% de la mediana nacional. El proceso de construcción ha requerido: (1) agregar los ingresos individuales de todos los miembros del hogar procedentes de diferentes fuentes (rentas del trabajo, prestaciones por desempleo, pensiones, transferencias sociales y otras fuentes recogidas en las variables PY010N a PY140N); (2) sumar los ingresos del hogar no atribuibles a miembros individuales (variables HY010 a HY145); (3) calcular la renta disponible total del hogar; (4) dividir por las unidades de consumo del hogar, calculadas según la escala de equivalencia de la OCDE modificada (1 unidad para el primer adulto, 0,5 para cada adulto adicional y 0,3 para cada menor de 14 años); y (5) calcular la mediana de esta renta equivalente para el conjunto de la muestra de cada año y fijar el umbral en el 60% de dicha mediana (Eurostat, 2025).

2.4.2. Componente de privación material y social severa

El segundo componente identifica a los hogares que no pueden permitirse al menos cinco de los trece ítems de la lista armonizada de Eurostat, actualizada en 2021 para incorporar la dimensión social junto a la estrictamente material. Los ítems incluyen aspectos como mantener la vivienda a temperatura adecuada, hacer frente a gastos imprevistos, disponer de determinados bienes duraderos (ordenador, vehículo, lavadora) y participar en actividades sociales y de ocio. Cada ítem se ha codificado como variable binaria (1 = privación, 0 = no privación) y se ha sumado el total de privaciones por hogar, definiendo la privación severa cuando este recuento es igual o superior a cinco (Eurostat, 2025).

2.4.3. Componente de baja intensidad laboral

El tercer componente mide la intensidad de trabajo de los miembros del hogar en edad laboral (18–59 años, excluidos los estudiantes de 18–24 años). El cálculo ha requerido: (1) identificar los meses de trabajo efectivo de cada miembro en edad laboral a partir de las variables mensuales de situación laboral (PL211A a PL211L); (2) sumar los meses trabajados de todos los miembros elegibles del hogar; (3) calcular los meses potenciales como el producto del número de miembros elegibles por doce; (4) calcular el cociente entre meses trabajados y meses potenciales; y (5) definir baja intensidad laboral cuando este cociente es inferior a 0,20 (Eurostat, 2025).

2.5. Construcción del indicador final y validación

Una vez contruidos los tres componentes, el indicador AROPE se ha definido como una variable binaria que toma valor 1 cuando el individuo cumple al menos una de las tres condiciones, y 0 en caso contrario. Esta operación de unión —no de suma— garantiza que cada individuo se contabiliza una sola vez independientemente del número de dimensiones en que se encuentre en situación de riesgo (Eurostat, 2025).

La validación del indicador construido se ha realizado comparando las tasas AROPE obtenidas para cada año con las publicadas oficialmente por el INE y Eurostat. Los

resultados muestran una coherencia elevada: las tasas calculadas a partir de los microdatos se sitúan dentro del margen de error muestral de las cifras oficiales en todos los años del período analizado, lo que confirma la robustez del proceso de construcción.

2.6. Muestra final y estructura del panel

La Tabla 2 resume las características de la muestra final utilizada en el análisis, distinguiendo entre la muestra completa (utilizada en el análisis descriptivo) y el panel equilibrado de cuatro años (utilizado en el análisis longitudinal).

Tabla 2. Caracterización de la muestra de análisis por año y submuestra. ECV España, 2020–2023.

Año	Observaciones individuales	Individuos únicos	Tasa AROPE (muestra completa)
2020	28.847	28.847	21,2%
2021	29.814	29.814	22,4%
2022	29.891	29.891	21,0%
2023	28.703	28.703	18,1%
Total (muestra completa)	117.255	48.618	19,38%
Panel equilibrado (4 años)	27.072	6.768	—

Nota: Elaboración propia a partir de los microdatos de la ECV (INE, 2020–2023). El panel equilibrado comprende los 6.768 individuos con observaciones válidas en los cuatro años consecutivos. La tasa AROPE de la muestra completa puede diferir ligeramente de las cifras oficiales por el tratamiento de los valores perdidos.

La muestra completa comprende 117.255 observaciones individuales correspondientes a 48.618 personas distintas. De ellas, 6.768 (13,9% del total de individuos únicos) cuentan con observaciones válidas en los cuatro años del panel y forman el panel equilibrado utilizado en el análisis de persistencia y transiciones. Esta submuestra no es representativa del conjunto de la población española, ya que excluye a los individuos que abandonan la encuesta o que se incorporan en años posteriores; sin embargo, es suficientemente amplia para el análisis de trayectorias y ofrece una imagen robusta de los patrones de persistencia del riesgo AROPE en el período estudiado.

2.7. Variables utilizadas en el análisis

La Tabla 3 presenta el conjunto de variables utilizadas en el análisis, con su descripción, tipo y papel en cada componente del trabajo.

Tabla 3. Variables utilizadas en el análisis descriptivo, longitudinal y predictivo.

Variable ECV	Descripción	Tipo	Uso en el análisis
AROPE	Indicador de riesgo de pobreza o exclusión social (construido)	Binaria	Variable dependiente
EDAD	Edad del individuo en años	Continua	Descriptivo + Modelo RF
PB150	Sexo (1=Hombre, 2=Mujer)	Categórica	Descriptivo + Modelo RF
PB190	Estado civil	Categórica	Descriptivo + Modelo RF
PE040 / PE041	Nivel de estudios más alto alcanzado (ISCED)	Ordinal	Descriptivo + Modelo RF
PL031	Situación de actividad laboral principal	Categórica	Descriptivo + Modelo RF

PL040	Situación en el empleo (asalariado/autónomo)	Categórica	Modelo RF
PL140	Tipo de contrato (indefinido/temporal)	Binaria	Modelo RF
PL145	Trabajo a tiempo parcial involuntario	Binaria	Modelo RF
TAMANO_HOGAR_GRUPO	Tamaño del hogar categorizado (1 / 2 / 3-4 / 5+)	Categórica	Descriptivo + Modelo RF
MENORES	Número de menores de 18 años en el hogar	Discreta	Descriptivo + Modelo RF
RENTA_UD_CONSUMO	Renta disponible equivalente por unidad de consumo	Continua	Análisis de renta (excluida del modelo RF)
PRIVACION_SEVERA	Privación material y social severa (componente AROPE)	Binaria	Componente AROPE (excluida del modelo RF)
PB030	Identificador personal longitudinal	Identificador	Análisis de panel
PB010	Año de referencia de la encuesta	Temporal	Análisis de panel

Nota: Elaboración propia a partir del diseño de registro de la ECV (INE, 2020–2023). RF = Random Forest. Las variables RENTA_UD_CONSUMO y PRIVACION_SEVERA se excluyen del modelo predictivo por ser componentes directos del AROPE, para evitar problemas de data leakage.

El nivel educativo (PE041) se ha categorizado en tres niveles siguiendo la clasificación ISCED: bajo (ISCED 0–2: sin estudios o educación primaria), medio (ISCED 3–4: educación secundaria) y alto (ISCED 5–8: educación superior). El tamaño del hogar se ha categorizado en cuatro grupos (1, 2, 3–4 y 5 o más personas). Para el análisis por nivel de renta del Capítulo III, se han definido tres grupos en función de la renta disponible equivalente por unidad de consumo: renta baja (≤ 7.859 €/año, equivalente a la media menos una desviación típica), renta alta (> 33.010 €/año, equivalente a la media más una

desviación típica) y renta media (el resto), siendo la media del período 20.435 € y la desviación típica 12.576 €.

2.8. Métodos de análisis

El trabajo combina tres aproximaciones metodológicas complementarias, cuya aplicación se detalla en los capítulos III y IV respectivamente.

Análisis descriptivo estático. Se calculan las tasas de incidencia del AROPE desagregadas por las principales variables sociodemográficas, utilizando la muestra completa de 117.255 observaciones. Las tablas de frecuencias y medias se han construido mediante tablas dinámicas en Excel y han sido verificadas con Python (pandas).

Análisis longitudinal de panel. A partir del panel equilibrado de 6.768 individuos, se calculan los patrones de persistencia del AROPE (número de años en situación de riesgo durante el período 2020–2023), las matrices de transición año a año (probabilidad condicional de cambio de estado entre dos años consecutivos) y el perfil sociodemográfico según el tipo de trayectoria. Este análisis sigue el marco conceptual de la literatura sobre pobreza dinámica (Ayllón, 2013), que distingue entre pobreza transitoria, intermitente y persistente en función de la duración y continuidad de los episodios de riesgo.

Modelo predictivo Random Forest. Se entrena un modelo de clasificación supervisada utilizando el algoritmo Random Forest (Breiman, 2001) sobre una submuestra de 7.590 individuos con observaciones completas en todas las variables del modelo. El modelo se implementa en Python con la librería scikit-learn (Pedregosa et al., 2011). La variable dependiente es el indicador AROPE; las variables predictoras son las características sociodemográficas y laborales listadas en la Tabla 3, excluyendo las componentes directas del AROPE para evitar *data leakage*. El modelo se configura con 500 árboles, pesos de clase balanceados y un umbral de decisión de 0,20, priorizando la sensibilidad (Recall) sobre la precisión (Precision) dado el contexto de detección de vulnerabilidad. Los detalles técnicos del modelo se presentan en el Capítulo IV.

CAPÍTULO III: ANÁLISIS DESCRIPTIVO Y LONGITUDINAL DEL AROPE EN ESPAÑA (2020–2023)

Este capítulo presenta el análisis empírico del riesgo de pobreza o exclusión social en España para el período 2020–2023. El análisis se estructura en dos bloques complementarios. En primer lugar, se realiza un análisis descriptivo estático que caracteriza la distribución del indicador AROPE según las principales variables sociodemográficas de la ECV. En segundo lugar, se explota la dimensión longitudinal de la encuesta para analizar la persistencia y las transiciones del riesgo a lo largo de los cuatro años del panel, que constituye el núcleo analítico del trabajo y su principal aportación respecto a los estudios de corte transversal existentes.

3.1. Construcción de la muestra de análisis

La base de datos utilizada en este capítulo se ha construido a partir de los microdatos de la Encuesta de Condiciones de Vida (ECV) para los años 2020, 2021, 2022 y 2023, tal y como se describe en el Capítulo II. La muestra total comprende 117.255 observaciones individuales correspondientes a 48.618 personas distintas. Dado que el objetivo principal del trabajo es analizar la dinámica longitudinal del AROPE, se trabaja con dos submuestras diferenciadas según el análisis.

Para el análisis descriptivo estático (apartado 3.2) se utiliza la muestra completa, lo que permite obtener tasas de incidencia representativas de todo el período. Para el análisis de persistencia y transiciones (apartado 3.3) se seleccionan los 6.768 individuos presentes en los cuatro años consecutivos del panel (panel equilibrado), lo que permite rastrear su trayectoria completa respecto al indicador AROPE.

En cuanto a la categorización de las variables sociodemográficas, el nivel educativo se ha clasificado a partir de la variable PE041 de la ECV siguiendo la clasificación ISCED: nivel bajo (sin estudios o educación primaria, ISCED 0–2), nivel medio (educación secundaria, ISCED 3–4) y nivel alto (educación superior, ISCED 5–8). El grupo de edad se ha construido en cuatro tramos (18–29, 30–44, 45–64 y 65 o más años) y el tamaño del

hogar en cuatro categorías (1, 2, 3–4 y 5 o más personas). Para el análisis por nivel de renta, se han definido tres grupos en función de la renta disponible equivalente por unidad de consumo: renta baja (por debajo de la media menos una desviación típica, ≤ 7.859 €/año), renta alta (por encima de la media más una desviación típica, > 33.010 €/año) y renta media (el resto), siendo la media del período 20.435 € y la desviación típica 12.576 €.

3.2. Análisis descriptivo estático: distribución del AROPE por características sociodemográficas

La Tabla 1 recoge las tasas de incidencia del AROPE para la población total y desglosadas por nivel de renta, según las principales variables sociodemográficas analizadas. La tasa media del período se sitúa en el 24,5%, con variaciones notables según las características individuales y del hogar.

Tabla 1. Tasas de riesgo de pobreza o exclusión social (ARPE) según características sociodemográficas y nivel de renta. España, 2020–2023.

Variable / Categoría	Población total (%)	Renta baja (≤ 7.859 €/año) (%)	Renta alta (> 33.010 €/año) (%)
Nivel educativo			
Alta	11,5%	100,0%	0,6%
Media	24,8%	100,0%	0,5%
Baja	25,0%	100,0%	1,1%
Sexo			

Hombre	23,2%	100,0%	1,0%
Mujer	25,8%	100,0%	1,1%
Grupo de edad			
18–29 años	30,6%	100,0%	2,5%
30–44 años	25,2%	100,0%	0,8%
45–64 años	23,9%	100,0%	1,0%
65 o más años	21,3%	100,0%	0,6%
Tamaño del hogar			
1 persona	32,2%	100,0%	1,6%
2 personas	21,3%	100,0%	0,8%
3–4 personas	22,0%	100,0%	1,0%
5 o más personas	40,1%	100,0%	2,1%
Total general	24,5%	100,0%	1,0%

Nota: Elaboración propia a partir de los microdatos de la ECV (INE, 2020–2023). El nivel educativo se clasifica según la escala ISCED a partir de la variable PE041: bajo (ISCED 0–2), medio (ISCED 3–4), alto (ISCED 5–8). Los grupos de renta se definen como: renta baja $\leq$ media – 1 SD ($\leq$ 7.859 €/año), renta alta $>$ media + 1 SD ($>$ 33.010 €/año), siendo la media 20.435 € y la SD 12.576 €. La tasa del 100% en el grupo de renta baja se explica porque este umbral se sitúa por debajo del umbral de pobreza monetaria del AROPE (60% de la mediana nacional), por lo que todos los individuos de este grupo cumplen al menos ese componente del indicador.

Nivel educativo. Se observa una relación clara e inversa entre el nivel educativo y el riesgo de pobreza. Los individuos con educación superior presentan una tasa del 11,5%, muy por debajo de quienes tienen educación media o baja (en torno al 25%). En el grupo de renta alta, esta diferencia se atenúa notablemente, con tasas inferiores al 1,1% en todos los niveles educativos, lo que indica que cuando los ingresos son elevados, el nivel educativo pierde capacidad discriminante respecto al AROPE. Este patrón es consistente con la evidencia sobre el papel del capital humano como factor protector frente a la exclusión social (Ayllón, 2013).

Sexo. Las mujeres presentan una tasa de AROPE ligeramente superior a la de los hombres (25,8% frente a 23,2%) en el conjunto de la población. Esta diferencia se mantiene en el grupo de renta alta (1,1% frente a 1,0%), aunque de forma muy reducida, lo que sugiere que la brecha de género en el riesgo de exclusión se concentra principalmente en los tramos medios y bajos de la distribución de ingresos (Eurostat, 2025).

Grupo de edad. Los jóvenes de 18 a 29 años presentan la tasa de AROPE más elevada (30,6%), significativamente por encima de los grupos de mayor edad. Esta mayor vulnerabilidad juvenil es especialmente acusada en el grupo de renta alta (2,5%), donde también presentan la tasa más elevada, lo que sugiere que la inestabilidad en los ingresos afecta a los jóvenes incluso en tramos de renta relativamente elevados, probablemente por la mayor prevalencia de contratos temporales y trayectorias laborales discontinuas al inicio de la vida profesional.

Tamaño del hogar. Se observa un patrón en forma de U: los hogares unipersonales (32,2%) y los de cinco o más miembros (40,1%) presentan las tasas más elevadas, mientras que los hogares de dos a cuatro personas se sitúan en torno al 21–22%. Esta dualidad refleja fenómenos distintos: en los hogares unipersonales, la ausencia de economías de escala y de ingresos compartidos incrementa la vulnerabilidad individual; en los hogares numerosos, el mayor número de dependientes presiona sobre los recursos disponibles del hogar. Esta diferencia se mantiene también en el grupo de renta alta, donde los hogares de cinco o más miembros (2,1%) y los unipersonales (1,6%) siguen siendo los más vulnerables (Eurostat, 2025).

3.3. Análisis longitudinal: persistencia y transiciones del AROPE (2020–2023)

El análisis longitudinal constituye la aportación diferencial de este trabajo respecto a los estudios de corte transversal sobre el AROPE en España. A partir de los 6.768 individuos del panel equilibrado de cuatro años, se analizan en primer lugar la evolución de la tasa AROPE año a año, en segundo lugar los patrones de persistencia y en tercer lugar las matrices de transición entre estados consecutivos.

3.3.1. Evolución del AROPE en el panel 2020–2023

La Tabla 2 muestra la evolución anual de la tasa AROPE entre los individuos del panel equilibrado.

Tabla 2. Evolución de la tasa AROPE en el panel equilibrado (N = 6.768). España, 2020–2023.

Año	Tasa AROPE (%)	Variación respecto al año anterior (pp)
2020	21,7%	—
2021	23,6%	+1,9
2022	21,7%	–1,9
2023	18,7%	–3,0
Media del período	21,4%	

Nota: Elaboración propia a partir de los microdatos de la ECV (INE, 2020–2023). Panel equilibrado de individuos observados en los cuatro años consecutivos.

La tasa AROPE alcanza su valor máximo en 2021 (23,6%), coincidiendo con el período de recuperación inmediata tras la crisis de la COVID-19, caracterizado por una elevada incertidumbre laboral y el agotamiento de las medidas de protección de emergencia adoptadas en 2020. A partir de 2022 se produce una reducción sostenida hasta situarse en el 18,7% en 2023, lo que refleja la mejora progresiva del mercado de trabajo en el contexto de la recuperación económica. Esta evolución es consistente con los datos publicados por el INE para el conjunto de la población española.

3.3.2. Patrones de persistencia

La Tabla 3 desglosa las trayectorias de los individuos del panel en función del número de años en que se encuentran en situación AROPE durante el período 2020–2023.

Tabla 3. Patrones de persistencia del AROPE en el panel equilibrado. España, 2020–2023.

Patrón de trayectoria	N individuos	% sobre panel 4 años
Nunca en AROPE (0,0,0,0)	4.456	65,8%
Siempre en AROPE — pobreza persistente (1,1,1,1)	673	9,9%
Trayectoria intermitente (total)	1.639	24,2%
· Solo 1 año en AROPE	670	9,9%
· 2 años en AROPE	471	7,0%
· 3 años en AROPE	498	7,4%
Total panel (N)	6.768	100%

Nota: Elaboración propia a partir de los microdatos de la ECV (INE, 2020–2023). Los porcentajes se calculan sobre el total del panel equilibrado (N = 6.768).

Los resultados revelan una marcada heterogeneidad en las trayectorias individuales respecto al AROPE. La mayoría de los individuos del panel (65,8%) no se encuentra en situación de riesgo en ninguno de los cuatro años analizados. Sin embargo, el 9,9% permanece en situación AROPE de forma ininterrumpida durante todo el período, lo que

constituye un indicador directo de pobreza persistente en sentido estricto. Este colectivo representa el núcleo duro de la exclusión social en España: personas cuya situación de vulnerabilidad no es transitoria ni atribuible a perturbaciones coyunturales, sino que responde a factores estructurales de largo plazo (Ayllón, 2013).

El 24,2% restante presenta trayectorias intermitentes, en las que los episodios de riesgo alternan con períodos fuera del umbral. De estos, el 9,9% experimenta únicamente un año en AROPE (pobreza transitoria), el 7,0% dos años y el 7,4% tres años. La distinción entre estas trayectorias tiene implicaciones directas para el diseño de políticas: los episodios transitorios de un solo año pueden responder a perturbaciones específicas susceptibles de atenuarse con medidas de activación laboral o rentas temporales de sustitución, mientras que los patrones de tres o cuatro años apuntan a una vulnerabilidad estructural que requiere intervenciones más intensivas y de mayor duración.

3.3.3. Matrices de transición año a año

Las matrices de transición permiten analizar la probabilidad de que un individuo cambie de estado —AROE o no AROE— entre dos años consecutivos. La Tabla 4 presenta estas probabilidades condicionadas para los tres períodos bienales del panel.

Tabla 4. Matrices de transición del AROPE entre años consecutivos. España, 2020–2023.

Período	No AROPE → No AROE	No AROPE → AROE (entrada)	AROE → No AROPE (salida)	AROE → AROE (persistencia)
2020→2021	90,9%	9,1%	23,8%	76,2%
2021→2022	93,5%	6,5%	29,1%	70,9%
2022→2023	93,6%	6,4%	36,9%	63,1%

Nota: Elaboración propia a partir de los microdatos de la ECV (INE, 2020–2023). Los porcentajes de cada fila suman 100% y representan la probabilidad condicional de encontrarse en cada estado en el año $t+1$ dado el estado en el año t .

Los resultados de las matrices de transición revelan tres patrones destacados. En primer lugar, existe una elevada estabilidad en el estado de no riesgo: entre el 90,9% y el 93,6% de los individuos que no están en situación AROPE en un año determinado tampoco lo están al año siguiente. En segundo lugar, la probabilidad de permanecer en situación AROPE de un año a otro —tasa de persistencia condicional— es muy elevada en todos los períodos, pero muestra una tendencia decreciente: del 76,2% en 2020–2021 al 70,9% en 2021–2022 y al 63,1% en 2022–2023, lo que refleja una mayor movilidad de salida del riesgo coincidiendo con la mejora del mercado de trabajo. En tercer lugar, la probabilidad de entrada al AROPE desde la situación de no riesgo también se reduce, del 9,1% en 2020–2021 al 6,4% en 2022–2023.

3.3.4. Perfil sociodemográfico según tipo de trayectoria

La Tabla 5 recoge el perfil sociodemográfico de los individuos del panel según el tipo de trayectoria AROPE experimentada.

Tabla 5. Perfil sociodemográfico según tipo de trayectoria AROPE. España, 2020–2023.

Variable	Nunca en AROPE	Intermitente	Siempre en AROPE (persistente)
Nivel educativo			
Alta	46,5%	24,6%	14,1%
Media	21,5%	25,0%	24,1%
Baja	32,0%	50,3%	61,8%
Sexo			
Hombre	49,6%	47,2%	44,9%
Mujer	50,4%	52,8%	55,1%
Grupo de edad			
18–29 años	10,0%	14,9%	13,4%
30–44 años	24,7%	22,8%	19,0%
45–64 años	42,3%	45,2%	44,0%

65 o más años	22,9%	17,0%	23,6%
---------------	-------	-------	-------

Nota: Elaboración propia a partir de los microdatos de la ECV (INE, 2020–2023). Los porcentajes de cada fila suman 100% y se calculan sobre el total de individuos de cada grupo de trayectoria. El perfil sociodemográfico se mide en el primer año de observación (2020).

En relación con el nivel educativo, los resultados son especialmente reveladores: el 61,8% de los individuos con pobreza persistente tiene un nivel educativo bajo, frente al 32,0% en el grupo que nunca entra en riesgo. En sentido contrario, solo el 14,1% de los individuos con pobreza persistente tiene educación superior, frente al 46,5% entre quienes nunca están en riesgo. El nivel educativo actúa, así, como el principal factor de protección frente a la cronicidad del riesgo (Ayllón, 2013).

En cuanto al sexo, se observa una feminización progresiva del riesgo a medida que aumenta su persistencia: las mujeres representan el 50,4% del grupo que nunca está en riesgo, el 52,8% del grupo intermitente y el 55,1% del grupo persistente. La tendencia es consistente con la mayor vulnerabilidad estructural femenina documentada en la literatura, aunque la diferencia no es extrema.

Por grupos de edad, los individuos de 45 a 64 años son el colectivo más numeroso en todos los tipos de trayectoria (en torno al 42–45%). Resulta relevante que los individuos de 65 o más años tienen una presencia mayor en el grupo persistente (23,6%) que en el intermitente (17,0%), lo que sugiere que la pobreza en edades avanzadas tiende a consolidarse en situaciones de largo plazo, probablemente vinculadas a pensiones de cuantía reducida.

3.4. Síntesis del análisis descriptivo y longitudinal

El análisis presentado en este capítulo permite establecer un conjunto de conclusiones que constituyen el fundamento empírico del modelo predictivo del Capítulo IV. El riesgo de pobreza o exclusión social en España no se distribuye de forma homogénea entre la población, sino que se concentra en colectivos con características específicas: jóvenes, personas con nivel educativo bajo, mujeres y hogares en los extremos del tamaño (unipersonales y de cinco o más miembros). Estas diferencias se mantienen al desagregar

por nivel de renta, aunque se atenúan notablemente en el grupo de renta alta, donde la tasa AROPE es inferior al 2,5% en todos los subgrupos analizados.

Más relevante aún, el análisis longitudinal revela que detrás de una tasa media de AROPE del 24,5% coexisten trayectorias muy heterogéneas. Un núcleo del 9,9% de los individuos del panel experimenta pobreza persistente durante los cuatro años analizados, mientras que casi una cuarta parte (24,2%) presenta episodios intermitentes de riesgo. Las matrices de transición confirman además que la probabilidad de permanecer en situación AROPE de un año al siguiente es muy elevada —entre el 63% y el 76% según el período—, aunque con una tendencia decreciente que refleja la mejora del mercado de trabajo a partir de 2022.

Estos resultados justifican el uso del modelo de Random Forest presentado en el Capítulo IV, cuyo objetivo es identificar con mayor precisión los factores que determinan tanto la entrada como la permanencia en situación de riesgo, aprovechando la capacidad de este algoritmo para capturar relaciones no lineales e interacciones entre variables que los métodos estadísticos tradicionales no detectan.

CAPÍTULO IV: MODELIZACIÓN PREDICTIVA MEDIANTE APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)

Este capítulo desarrolla el componente predictivo del trabajo. Partiendo de los patrones de riesgo identificados en el análisis descriptivo y longitudinal del Capítulo III, se construye un modelo de aprendizaje automático cuyo objetivo es identificar los factores que con mayor fuerza determinan el riesgo de pobreza o exclusión social a nivel individual. El capítulo se estructura en cuatro apartados: justificación de la aproximación algorítmica, fundamentos teóricos del Random Forest, especificación técnica del modelo y evaluación de resultados.

4.1. Justificación de la aproximación algorítmica

El análisis descriptivo del Capítulo III ha puesto de manifiesto que el riesgo AROPE responde a una estructura multicausal en la que interactúan de forma simultánea factores demográficos, educativos, laborales y del hogar. Esta complejidad plantea un reto metodológico para los enfoques estadísticos tradicionales: los modelos de regresión logística, aunque habituales en la literatura sobre pobreza, asumen relaciones lineales entre predictores y variable dependiente, e imponen restricciones sobre las interacciones entre variables que pueden sesgar la estimación cuando las relaciones son de naturaleza no lineal (Breiman, 2001).

El algoritmo de Random Forest se presenta como una alternativa metodológica adecuada para este tipo de problemas por varias razones. En primer lugar, es capaz de capturar relaciones no lineales e interacciones de alto orden entre predictores sin necesidad de especificarlas a priori. En segundo lugar, es robusto frente al ruido en los datos y al sobreajuste, característica especialmente relevante cuando se trabaja con microdatos de encuesta como la ECV. En tercer lugar, proporciona una medida de importancia de variables (*feature importance*) que permite jerarquizar los factores predictores de forma transparente e interpretable, lo que resulta especialmente valioso para extraer conclusiones de política pública (Hastie, Tibshirani y Friedman, 2009).

Desde el punto de vista del objetivo del trabajo, la modelización predictiva cumple una función complementaria al análisis longitudinal: mientras que el Capítulo III describe quién está en situación AROPE y durante cuánto tiempo, el modelo del presente capítulo permite identificar qué características individuales y del hogar predicen con mayor fuerza esa situación, proporcionando así una base científica para el diseño de sistemas de detección temprana y políticas de prevención.

4.2. Fundamentos del algoritmo Random Forest

El Random Forest es un método de aprendizaje automático supervisado basado en el principio de combinación de modelos (*ensemble learning*). Su lógica fundamental consiste en construir un número elevado de árboles de decisión independientes, cada uno entrenado sobre una submuestra aleatoria de los datos y utilizando en cada nodo un subconjunto aleatorio de predictores. La predicción final del modelo es el resultado de agregar las predicciones individuales de todos los árboles —por votación mayoritaria en problemas de clasificación—, lo que reduce la varianza del estimador y mejora la capacidad de generalización respecto a un árbol de decisión único (Breiman, 2001).

Cada árbol del bosque se construye mediante el algoritmo CART (*Classification and Regression Trees*), que divide recursivamente el espacio de predictores en regiones que maximizan la homogeneidad de la variable dependiente dentro de cada nodo. La aleatorización en la selección de variables en cada división —conocida como *feature bagging*— introduce diversidad entre los árboles y reduce la correlación entre ellos, lo que es la principal fuente de la superioridad del conjunto sobre el árbol individual (Hastie et al., 2009).

La importancia de cada variable se calcula como la reducción media del índice de impureza de Gini que aporta esa variable a lo largo de todos los árboles del bosque. Variables con mayor importancia son, en promedio, las que producen particiones más limpias entre individuos en riesgo y no en riesgo, y por tanto las que más contribuyen a la capacidad predictiva del modelo (Breiman, 2001).

4.3. Especificación técnica y entrenamiento del modelo

El modelo se ha especificado e implementado en Python (v3.12) utilizando la librería scikit-learn (Pedregosa et al., 2011). La variable dependiente es el indicador AROPE (binario: 1 = en riesgo, 0 = no en riesgo). Las variables predictoras incluyen características individuales y del hogar disponibles en la ECV que no forman parte directa del cálculo del indicador, con el fin de evitar problemas de data leakage —esto es, la inclusión de información que el modelo no podría conocer en un contexto de detección real—. Concretamente, se han excluido la renta disponible por unidad de consumo y la privación material severa, que son componentes directos del AROPE, y se ha trabajado exclusivamente con características sociodemográficas y laborales: edad, sexo, estado civil, nivel educativo (PE040), situación de actividad (PL031), situación laboral (PL040), tipo de contrato (PL140), trabajo a tiempo parcial involuntario (PL145), tamaño del hogar y número de menores en el hogar.

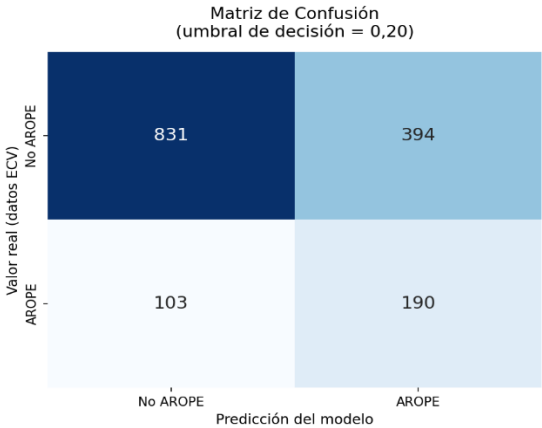
La muestra final, tras la eliminación de observaciones con valores perdidos en alguna de las variables del modelo, comprende 7.590 individuos. El conjunto de datos se ha dividido aleatoriamente en un 80% para entrenamiento (6.072 observaciones) y un 20% para evaluación (1.518 observaciones), fijando una semilla aleatoria (`random_state = 42`) para garantizar la reproducibilidad de los resultados.

El modelo se ha configurado con 500 árboles de decisión (`n_estimators = 500`). Dado el desequilibrio de clases en la muestra —la prevalencia del AROPE en la muestra de test es del 19,3%—, se ha aplicado la técnica de pesos de clase balanceados (`class_weight = 'balanced'`), que asigna a cada observación de la clase minoritaria (AROEPE = 1) un peso inversamente proporcional a su frecuencia, compensando así el sesgo del modelo hacia la clase mayoritaria. Adicionalmente, se ha ajustado el umbral de decisión probabilístico a 0,20 en lugar del estándar de 0,50, priorizando la sensibilidad del modelo (Recall) sobre su precisión (Precision). Esta elección se justifica por el contexto de aplicación: en un sistema de detección temprana de pobreza, el coste de no detectar a un individuo en riesgo real (falso negativo) es socialmente más elevado que el de generar una alerta sobre un individuo que en realidad no está en riesgo (falso positivo) (Hastie et al., 2009).

4.4. Evaluación del rendimiento y capacidad de detección

La evaluación del modelo se realiza sobre la muestra de test ($N = 1.518$), no vista por el algoritmo durante el entrenamiento. La Figura 1 presenta la matriz de confusión, que cruza los valores reales de la variable AROPE con las predicciones del modelo.

Figura 1. Matriz de confusión del modelo Random Forest. España, ECV 2020–2023.



Nota: Elaboración propia a partir de los microdatos de la ECV (INE, 2020–2023). Umbral de decisión = 0,20. $N_{test} = 1.518$ observaciones.

La Tabla 1 recoge las métricas de evaluación del modelo.

Tabla 1. Métricas de evaluación del modelo Random Forest. España, ECV 2020–2023.

Métrica	Definición	Valor obtenido	Interpretación
Accuracy (exactitud)	% de clasificaciones correctas sobre el total	67,3%	El modelo clasifica correctamente 2 de cada 3 observaciones
Recall / Sensibilidad	% de casos AROPE reales que el modelo detecta	64,8%	Detecta correctamente al 64,8% de los individuos realmente en riesgo
Precision	% de casos predichos como AROPE que lo son realmente	32,5%	1 de cada 3 alertas generadas corresponde a un caso real de riesgo
F1-Score	Media armónica entre Recall y Precision	0,433	Equilibrio moderado entre detección y precisión

Nota: Elaboración propia. Las métricas se calculan sobre la muestra de test ($N = 1.518$). Umbral de decisión = 0,20.

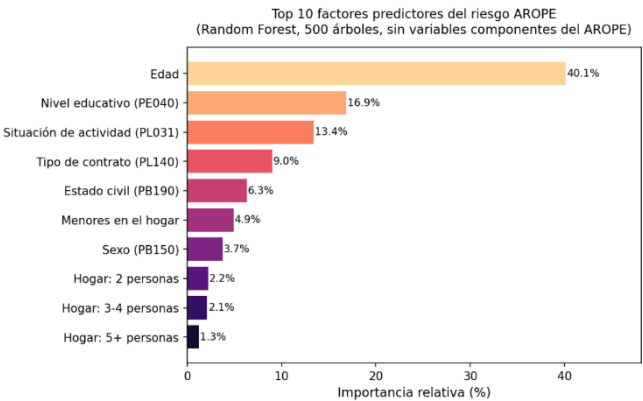
El modelo alcanza una tasa de detección (Recall) del 64,8%, lo que significa que identifica correctamente a 190 de los 293 individuos realmente en situación AROPE en la muestra de test. Los 103 casos no detectados (falsos negativos) representan el principal límite del modelo cuando se trabaja únicamente con variables sociodemográficas y laborales, sin incluir información directa sobre ingresos o privación. Este resultado es esperable y metodológicamente coherente: la exclusión de las variables componentes del AROPE hace el problema de clasificación considerablemente más difícil, pero también más realista desde el punto de vista de la detección temprana.

La exactitud global (Accuracy) del 67,3% y el F1-Score de 0,433 reflejan un equilibrio moderado entre la capacidad de detección y la tasa de falsas alarmas. En particular, la precisión del 32,5% indica que aproximadamente 1 de cada 3 alertas generadas por el modelo corresponde a un individuo realmente en riesgo. En el contexto de política social, este nivel de falsas alarmas es gestionable si el modelo se utiliza como herramienta de cribado inicial —orientando recursos de diagnóstico más intensivo hacia la población señalada— y no como instrumento de asignación automática de prestaciones.

4.5. Jerarquía de determinantes del riesgo AROPE

Más allá de la capacidad predictiva global, el Random Forest permite extraer la importancia relativa de cada variable en la determinación del riesgo AROPE. La Figura 2 y la Tabla 2 presentan los diez factores con mayor peso predictivo en el modelo.

Figura 2. Importancia relativa de las variables predictoras del riesgo AROPE. España, ECV 2020–2023.



Nota: Elaboración propia a partir de los microdatos de la ECV (INE, 2020–2023). La importancia se mide como la reducción media del índice de Gini a lo largo de los 500 árboles del modelo.

Tabla 2. Importancia relativa de las diez principales variables predictoras del riesgo AROPE. España, ECV 2020–2023.

Posición	Variable	Descripción	Importancia relativa
1º	EDAD	Edad del individuo	40,1%
2º	PE040	Nivel educativo más alto alcanzado	16,9%
3º	PL031	Situación de actividad laboral	13,4%
4º	PL140	Tipo de contrato (indefinido/temporal)	9,0%
5º	PB190	Estado civil	6,3%
6º	MENORES	Número de menores en el hogar	4,9%
7º	PB150	Sexo	3,7%
8º	TAMANO_HOGAR_GRUPO_2	Hogar de 2 personas	2,2%
9º	TAMANO_HOGAR_GRUPO_3-4	Hogar de 3-4 personas	2,1%
10º	TAMANO_HOGAR_GRUPO_5+	Hogar de 5 o más personas	1,3%

Nota: Elaboración propia. Los porcentajes suman 100% sobre el conjunto de variables del modelo.

Los resultados revelan una estructura de determinantes con tres niveles de relevancia claramente diferenciados. En primer lugar, la edad emerge como el predictor más potente con una importancia relativa del 40,1%, lo que refleja que el ciclo vital —y con él la trayectoria laboral acumulada, los derechos sociales adquiridos y la situación familiar— es el factor que con mayor fuerza condiciona el riesgo de exclusión a nivel individual. Este resultado es consistente con el hallazgo del Capítulo III, donde los jóvenes de 18 a 29 años presentaban la tasa AROPE más elevada (30,6%).

En segundo nivel de relevancia se sitúan el nivel educativo (PE040, 16,9%) y la situación de actividad laboral (PL031, 13,4%), que en conjunto aportan el 30,3% de la capacidad predictiva del modelo. Estos resultados confirman que el capital humano y la inserción en el mercado de trabajo constituyen los principales vectores de protección frente a la exclusión social: la educación no solo determina el nivel de ingresos potenciales, sino también la estabilidad y la calidad del empleo, mientras que la situación de actividad refleja directamente la participación en los circuitos formales de protección social. La literatura sobre persistencia de la pobreza en España señala de forma consistente estos dos factores como los determinantes estructurales más relevantes de las trayectorias de largo plazo en situación de riesgo (Ayllón, 2013).

En tercer nivel, el tipo de contrato (PL140, 9,0%) y el estado civil (PB190, 6,3%) completan el núcleo explicativo del modelo. La relevancia del tipo de contrato confirma que la precariedad laboral —medida por la temporalidad— actúa como factor de riesgo independiente de la situación de actividad: incluso entre los ocupados, quienes tienen contratos temporales presentan una probabilidad significativamente mayor de estar en situación AROPE. El estado civil, por su parte, captura efectos de economías de escala y de redes de apoyo intra-hogar que amortiguan o amplifican el riesgo individual.

Finalmente, el número de menores en el hogar (4,9%) y el sexo (3,7%) se posicionan como factores de segunda línea, cuya importancia es menor pero estadísticamente relevante. La presencia de menores dependientes incrementa las necesidades del hogar sin un incremento proporcional de los ingresos, mientras que el sexo recoge la desventaja estructural femenina documentada en el análisis descriptivo. Estos resultados son coherentes con la feminización progresiva de la pobreza persistente observada en el Capítulo III.

4.6. Síntesis del capítulo

La aplicación del algoritmo Random Forest a los microdatos de la ECV ha permitido construir un modelo predictivo del riesgo AROPE que, trabajando exclusivamente con características sociodemográficas y laborales observables —sin recurrir a las variables que componen directamente el indicador—, alcanza una tasa de detección del 64,8% con una exactitud global del 67,3%. Aunque estos valores son moderados, son metodológicamente robustos y representan una estimación realista de la capacidad de detección temprana a partir de información fácilmente disponible en registros administrativos o encuestas básicas.

Los resultados del análisis de importancia de variables ofrecen, además, una jerarquía clara de los determinantes del riesgo: la edad, el nivel educativo y la situación de actividad laboral concentran el 70% de la capacidad predictiva del modelo, lo que subraya el papel central del ciclo vital y del capital humano en la configuración del riesgo de exclusión. La temporalidad contractual y la presencia de menores en el hogar completan el perfil de riesgo con una contribución estadísticamente relevante. Estos hallazgos son consistentes con los patrones identificados en el análisis longitudinal del Capítulo III y proporcionan una base empírica sólida para las recomendaciones de política pública que se desarrollan en el Capítulo V.

CAPÍTULO V: CONCLUSIONES

Este Trabajo Fin de Grado ha abordado la persistencia y las transiciones del riesgo de pobreza o exclusión social en España a partir de los microdatos de la Encuesta de Condiciones de Vida (ECV/EU-SILC) para el período 2020–2023. El trabajo ha combinado un análisis descriptivo estático, un análisis longitudinal de panel y un modelo predictivo de aprendizaje automático, con el objetivo de ofrecer una visión más completa del fenómeno que la que proporcionan los estudios de corte transversal. Las conclusiones se organizan en torno a los tres ejes analíticos del trabajo.

5.1 Sobre la distribución y los determinantes estáticos del riesgo AROPE

El análisis descriptivo confirma que el riesgo de pobreza o exclusión social en España no se distribuye de forma homogénea entre la población: la tasa media del período 2020–2023 se sitúa en el 24,5%, pero con variaciones muy significativas según las características individuales y del hogar. Los resultados permiten extraer cuatro conclusiones de carácter transversal.

Primera. El nivel educativo es el factor de protección más potente frente al riesgo AROPE en el conjunto del período analizado. Los individuos con educación superior presentan una tasa de riesgo del 11,5%, frente al 25% de quienes tienen educación media o baja. Esta diferencia, que se mantiene al desagregar por nivel de renta, confirma el papel central del capital humano como vector de inserción laboral estable y de acceso a rentas suficientes para superar los umbrales del indicador (Ayllón, 2013).

Segunda. Los jóvenes de 18 a 29 años presentan la mayor vulnerabilidad entre todos los grupos de edad (30,6%), una diferencia que persiste incluso en el grupo de renta alta, lo que sugiere que la precariedad laboral juvenil —contratos temporales, bajas remuneraciones iniciales, trayectorias discontinuas— genera una exposición al riesgo que no se resuelve simplemente con ingresos más elevados. Este hallazgo apunta a la necesidad de políticas que actúen específicamente sobre la calidad del empleo juvenil, no solo sobre su cantidad.

Tercera. La estructura del hogar introduce una dualidad de riesgo que merece atención diferenciada. Los hogares unipersonales (27,2%) y los de cinco o más miembros (40,1%) presentan tasas muy superiores a los hogares de dos a cuatro personas (en torno al 17–22%). Estas dos situaciones responden a lógicas distintas: en los hogares unipersonales el riesgo deriva de la ausencia de economías de escala y de ingresos compartidos; en los hogares numerosos, de la mayor presión de las necesidades sobre recursos que no crecen proporcionalmente.

Cuarta. Las mujeres presentan una tasa de AROPE superior a la de los hombres (25,8% frente a 23,2%), diferencia que aumenta progresivamente en los colectivos con mayor persistencia del riesgo. Aunque la brecha no es extrema en términos estáticos, su acentuación en el análisis longitudinal apunta a una desventaja estructural acumulativa que los datos transversales tienden a subestimar.

5.2 Sobre la persistencia y las transiciones del riesgo AROPE

El análisis longitudinal del panel de 6.768 individuos observados durante los cuatro años del período constituye la aportación diferencial de este trabajo respecto a los estudios existentes sobre el AROPE en España. Los resultados permiten formular tres conclusiones adicionales.

Quinta. El 9,9% de los individuos del panel permanece en situación AROPE de forma ininterrumpida durante los cuatro años analizados. Este colectivo —que puede calificarse como pobreza persistente en sentido estricto— no responde a perturbaciones coyunturales sino a factores estructurales profundos. Su perfil es claro: predominio de nivel educativo bajo (61,8%), sobrerrepresentación de mujeres (55,1%) y mayor presencia de personas de 65 o más años (23,6%) en comparación con los grupos de trayectoria intermitente. La pobreza persistente es, por definición, la que más difícilmente se resuelve con medidas de activación laboral a corto plazo, y la que requiere intervenciones más intensivas y sostenidas en el tiempo.

Sexta. Las matrices de transición revelan una elevada inercia en el estado de riesgo: la probabilidad de permanecer en situación AROPE de un año al siguiente oscila entre el 76,2% (2020–2021) y el 63,1% (2022–2023). Esta inercia decreciente a lo largo del

período es una señal positiva —refleja la mejora del mercado de trabajo a partir de 2022— pero también subraya que, incluso en el mejor de los escenarios del período analizado, más de 6 de cada 10 individuos en riesgo siguen estándolo al año siguiente. La salida del AROPE no es un fenómeno espontáneo: requiere cambios estructurales en la situación laboral, educativa o familiar del individuo.

Séptima. El 24,2% de los individuos del panel presenta trayectorias intermitentes, con episodios de riesgo que alternan con períodos fuera del umbral. La distinción entre este grupo y el persistente tiene implicaciones directas para el diseño de políticas: los episodios de un solo año (9,9% de la muestra) responden mejor a medidas de activación y rentas temporales de sustitución, mientras que las trayectorias de tres o cuatro años requieren intervenciones más estructurales. Agrupar ambas situaciones bajo la misma categoría estadística —como hace el AROPE en su versión estática anual— oculta una heterogeneidad que es crucial para la eficiencia de la política social.

5.3 Sobre el modelo predictivo y su potencial para la detección temprana

Octava. El modelo Random Forest construido con variables sociodemográficas y laborales —excluyendo deliberadamente las variables componentes del AROPE para evitar problemas de *data leakage*— alcanza una tasa de detección (Recall) del 64,8% con una exactitud global del 67,3%. Estos valores son moderados, pero metodológicamente robustos: representan la capacidad real de detección a partir de información observable en registros administrativos o encuestas básicas, sin necesidad de conocer los ingresos o la privación material del individuo. Este resultado tiene un valor práctico relevante para el diseño de sistemas de cribado preventivo.

Novena. La jerarquía de determinantes obtenida mediante el análisis de importancia de variables confirma y enriquece los hallazgos del análisis descriptivo. La edad (40,1%), el nivel educativo (16,9%) y la situación de actividad laboral (13,4%) concentran el 70% de la capacidad predictiva del modelo, lo que subraya el papel del ciclo vital y del capital humano como ejes estructuradores del riesgo. El tipo de contrato (9,0%) añade un factor

de calidad del empleo que los análisis de actividad simplemente activo/inactivo no capturan. La presencia de menores en el hogar (4,9%) y el sexo (3,7%) completan el perfil de riesgo con factores de vulnerabilidad familiar y de género que refuerzan los resultados del análisis longitudinal.

5.4 Limitaciones del trabajo y líneas de investigación futura

Este trabajo presenta limitaciones que deben tenerse en cuenta en la interpretación de los resultados. En primer lugar, el panel equilibrado utilizado en el análisis longitudinal (N = 6.768) excluye por definición a los individuos con observaciones incompletas, lo que puede introducir un sesgo de selección si los individuos que abandonan la muestra presentan características sistemáticamente distintas. En segundo lugar, el modelo predictivo opera a nivel de individuo-año sin incorporar la dimensión temporal de las trayectorias, lo que limita su capacidad para distinguir entre perfiles de pobreza transitoria y persistente en el momento de la predicción.

Estas limitaciones abren líneas de investigación relevantes. La aplicación de modelos de secuencia de estados (*sequence analysis*) o de modelos de duración a los datos de panel de la ECV permitiría analizar con mayor precisión los factores que determinan la longitud de los episodios de riesgo. Asimismo, la incorporación de variables territoriales o de acceso a servicios públicos podría mejorar la capacidad predictiva del modelo y añadir una dimensión de política local a los resultados.

REFERENCIAS

1. Atkinson, A. B., Cantillon, B., Marlier, E. y Nolan, B. (2002). *Social Indicators: The EU and Social Inclusion*. Oxford University Press.
2. Atkinson, A. B., Marlier, E. y Nolan, B. (2004). Indicators and Targets for Social Inclusion in the European Union. *Journal of Common Market Studies*, 42(1), 47–75.
3. Ayllón, S. (2013). Understanding poverty persistence in Spain. *SERIEs – Journal of the Spanish Economic Association*, 4(2), 201–233. <https://doi.org/10.1007/s13209-012-0089-4>
4. Ayllón, S. (2013). Understanding poverty persistence in Spain. *SERIEs – Journal of the Spanish Economic Association*, 4(2), 201–233. <https://doi.org/10.1007/s13209-012-0089-4>
5. Ayllón, S. (2013). Understanding poverty persistence in Spain. *SERIEs – Journal of the Spanish Economic Association*, 4(2), 201–233. <https://doi.org/10.1007/s13209-012-0089-4>
6. Ayllón, S. (2013). Understanding poverty persistence in Spain. *SERIEs – Journal of the Spanish Economic Association*, 4(2), 201–233. <https://doi.org/10.1007/s13209-012-0089-4>
7. Ayllón, S. (2013). Understanding poverty persistence in Spain. *SERIEs – Journal of the Spanish Economic Association*, 4(2), 201–233. <https://doi.org/10.1007/s13209-012-0089-4>
8. Banco de España (2018). *La desigualdad de la renta, el consumo y la riqueza en España*. Banco de España.
9. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
10. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
11. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
12. EAPN-ES — Red Europea de Lucha contra la Pobreza y la Exclusión Social en el Estado Español (2023). *Informe El Estado de la Pobreza 2023 – Comunidad de Madrid*. EAPN-ES, Madrid.

13. Eurostat (2025). EU statistics on income and living conditions (EU-SILC) methodology – People at risk of poverty or social exclusion. Statistics Explained. Comisión Europea. <https://ec.europa.eu/eurostat/statistics-explained>
14. Eurostat (2025). EU statistics on income and living conditions (EU-SILC) methodology – People at risk of poverty or social exclusion. Statistics Explained. Comisión Europea. [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_statistics_on_income_and_living_conditions_\(EU-SILC\)_methodology_-_people_at_risk_of_poverty_or_social_exclusion](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology_-_people_at_risk_of_poverty_or_social_exclusion)
15. Eurostat (2025). EU statistics on income and living conditions (EU-SILC) methodology – Introduction. Statistics Explained. Comisión Europea. [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_statistics_on_income_and_living_conditions_\(EU-SILC\)_methodology_-_introduction](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology_-_introduction)
16. Eurostat (2025). EU statistics on income and living conditions (EU-SILC) methodology – Introduction. Statistics Explained. Comisión Europea. <https://ec.europa.eu/eurostat/statistics-explained>
17. Eurostat (2025). EU statistics on income and living conditions (EU-SILC) methodology – People at risk of poverty or social exclusion. Statistics Explained. Comisión Europea. <https://ec.europa.eu/eurostat/statistics-explained>
18. Eurostat (2025). EU statistics on income and living conditions (EU-SILC) methodology – People at risk of poverty or social exclusion. Statistics Explained. Comisión Europea. <https://ec.europa.eu/eurostat/statistics-explained>
19. García-Pardo, I., Bárcena-Martín, E. y Pérez-Moreno, S. (2020). Measuring the 'leaving no one behind' principle in the European countries: An AROPE-based fuzzy logic approach. Social Indicators Research.
20. Hastie, T., Tibshirani, R. y Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2.^a ed.). Springer.
21. Instituto Nacional de Estadística (INE) (2020–2023). Encuesta de Condiciones de Vida (ECV). Microdatos y documentación metodológica. <https://www.ine.es>
22. Instituto Nacional de Estadística (INE) (2020–2023). Encuesta de Condiciones de Vida (ECV). Microdatos. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176807&menu=resultados&idp=1254735976608

23. Instituto Nacional de Estadística (INE) (2020–2023). Encuesta de Condiciones de Vida (ECV). Microdatos. <https://www.ine.es>
24. Instituto Nacional de Estadística (INE) (2020–2023). Encuesta de Condiciones de Vida (ECV). Microdatos. <https://www.ine.es>
25. Instituto Nacional de Estadística (INE) (2025). Mujeres y hombres en España 2025. Apartado 2.11: Riesgo de pobreza o exclusión social (objetivo Europa 2030). Indicador AROPE. INE. <https://www.ine.es>
26. Lafuente Lechuga, M., García Luque, O. y Faura Martínez, U. (2015). Crisis y evolución regional del indicador AROPE en España. *Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA. Rect@*, 16, 91–104.
27. Llano Ortiz, J. C. (2019). El Estado de la Pobreza. Seguimiento del indicador de pobreza y exclusión social en España 2008–2018. 9.º Informe. EAPN-ES, Madrid.
28. Ministerio de Derechos Sociales, Consumo y Agenda 2030 (2024). Informe de indicadores 2019–2024. https://www.dsca.gob.es/sites/default/files/derechossociales/inclusion/docs/INFORME_INDICADORES_2019-2024_DEF.pdf
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... y Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
30. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... y Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
31. Podadera Rivera, P., Calderón Milán, M. J. y García Perea, P. (2020). Pobreza y exclusión social en la Unión Europea. [Editorial/Revista pendiente de verificar con la versión completa de la referencia].
32. Room, G. (1995). Poverty and social exclusion: The new European agenda for policy and research. En G. Room (Ed.), *Beyond the Threshold: The Measurement and Analysis of Social Exclusion* (pp. 1–9).
33. Sen, A. (1983). Poor, relatively speaking. *Oxford Economic Papers*, 35(2), 153–169.

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

Por la presente, yo, María Caballero Pascual, estudiante de E-3 Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado “PERSISTENCIA Y TRANSICIONES DEL RIESGO DE POBREZA O EXCLUSIÓN SOCIAL EN ESPAÑA: EVIDENCIA CON DATOS DE PANEL ECV (EU-SILC)”, 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. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
 8. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
 9. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
 10. **Generador de datos sintéticos de prueba:** Para la creación de conjuntos de datos ficticios.
 11. **Generador de problemas de ejemplo:** Para ilustrar conceptos y técnicas.
- Traductor:** Para traducir textos de un lenguaje a otro.

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: 1/06/2026

Firma: María Caballero Pascual