Self-employment and Machine Learning: An application for Spain.

Gutierrez-Lythgoe, Antonio

Universidad de Zaragoza

May 2023

Autoempleo y Machine Learning: Una aplicación para España

Antonio Gutiérrez-Lythgoe

Mayo, 2023

Resumen

La investigación en el campo de la Inteligencia Artificial ha avanzado considerablemente en los últimos años, demostrando su eficacia en la predicción y clasificación de decisiones discretas. Sin embargo, estos avances han sido relativamente poco empleados en la investigación en economía debido a la falta de vínculos con teorías económicas que expliquen el proceso de toma de decisiones de los agentes. En este trabajo, proponemos un marco microeconómico para el árbol de decisión, una técnica de Aprendizaje Automático, para establecer una conexión más sólida con la teoría económica y fomentar su aplicación en el campo de la elección discreta. Para ello, recurrimos a los datos de la EU-SILC del año 2019 para España. A través de la comparación con un modelo logit multinomial convencional, demostramos la utilidad de esta perspectiva económica para el estudio de los factores sociodemográficos asociados al autoempleo en España. Los resultados sugieren que la incorporación de fundamentos económicos puede mejorar significativamente la precisión de las predicciones y la capacidad de dibujar perfiles sociodemográficos individuales para el autoempleo.

Abstract

Research in the field of Artificial Intelligence has made considerable progress in recent years, demonstrating its effectiveness in predicting and classifying discrete decisions. However, these advances have been relatively underutilized in economic research due to the lack of links with economic theories that explain the decision-making process of agents. In this paper, we propose a microeconomic framework for decision trees, a machine learning technique, to establish a more solid connection with economic theory and encourage its application in the field of discrete choice. To do so, we rely on data from the 2019 EU-SILC for Spain. Through comparison with a conventional multinomial logit model, we demonstrate the usefulness of this economic perspective for studying the sociodemographic factors associated with self-employment in Spain. The results suggest that incorporating economic foundations can significantly improve the accuracy of predictions and the ability to draw individual sociodemographic profiles for self-employment.

Keywords: Artificial Intelligence, Machine Learning, Microeconomics, Self-employment, multinomial logit

JEL Classification: C45, C53, J24, J62, L26

Introducción

En los últimos tiempos, ha aumentado el interés en la Inteligencia Artificial (IA), impulsado por importantes avances, como en el campo de la IA generativa con el lanzamiento de aplicaciones como Chat-GPT. Estos avances han puesto de manifiesto el gran potencial de la IA para transformar el mercado de trabajo y la actividad laboral (Acemoglu & Restrepo, 2020; Eloundou et al., 2023) y la convierten en una herramienta atractiva para su uso en distintos campos de investigación (Boyd & Crawford, 2012; Einav & Levin, 2014). La capacidad de la IA para procesar grandes cantidades de datos y generar patrones y predicciones precisas promete avances en el estudio de distintas áreas de las ciencias sociales, como la economía (Einav & Levin, 2014), la política económica (Agrawal et al., 2019; Lambert & Fegley, 2023) o la psicología (Min et al., 2021). En este contexto, el estudio del emprendimiento también puede beneficiarse de los nuevos enfoques que ofrece la IA y el análisis de Big Data (Obschonka & Audretsch, 2020). Por lo tanto, en esta investigación, se implementaron técnicas de IA para comprender las decisiones de los individuos en el mercado laboral, en particular, utilizando técnicas de aprendizaje automático para la modelización discreta, como los árboles de decisión.

Debemos apreciar la diferencia metodológica que esto supone con los enfoques utilizados tradicionalmente en la literatura empírica en economía. La diferencia entre la aproximación mediante algoritmos y la modelización estadística (o econométrica, en el caso de la economía) es significativa (Breiman, 2001; Athey & Imbens, 2019). Según Breiman (2001), la modelización estadística asume que los datos son generados por un modelo de datos estocástico dado. En cambio, la aproximación mediante algoritmos trata al mecanismo de los datos como desconocido. En los últimos años se han realizado contribuciones en la comunidad estadística con el objetivo de incorporar este tipo de metodologías en la modelización (Athey & Imbens, 2019). Sin embargo, en el campo de la investigación en economía, esto ha sucedido en menor magnitud, posiblemente porque se prefieren métodos que tengan propiedades formales, como la consistencia, normalidad y eficiencia, las cuales no son necesariamente proporcionadas por los métodos de ML (Athey & Imbens, 2019). No obstante, es necesario profundizar e investigar sobre estas metodologías, que en muchos casos pueden mejorar la precisión de la modelización económica (Brathwaite et al., 2017; Athey & Imbens, 2019).

En particular, para modelar las decisiones de los individuos en el mercado laboral se puede recurrir a la literatura sobre modelización discreta (Einhorn, 1970; Tversky, 1972; McFadden, 1973; Hauser et al., 2010; Brathwaite et al., 2017). Los modelos de decisión discreta se utilizan en muchas aplicaciones microeconómicas con el objetivo de comprender los razonamientos económicos que motivan las decisiones de los individuos (McFadden, 1973; Brathwaite et al., 2017). De esta manera, estos modelos logran mostrar con estimaciones insesgadas cómo unas determinadas preferencias afectan las decisiones de los individuos. Este tipo de modelos se ha utilizado ampliamente en el contexto de la movilidad urbana y elección de modos de transporte (McFadden, 1973; Brathwaite et al., 2017; Paredes et al., 2017).

El modelo clásico de elección discreta es el modelo propuesto por McFadden (1973), el modelo multinomial logit. Es una contribución muy relevante en la literatura porque logra vincular un modelo de elección discreta con la teoría de la decisión. En este artículo,

proponemos utilizar algoritmos propios de Inteligencia Artificial, concretamente de Aprendizaje Automático (Machine Learning), para modelar elecciones discretas. La realidad es que este tipo de modelos puede mejorar la precisión de los modelos econométricos convencionales, pero a expensas de perder la explicabilidad o el razonamiento económico detrás de estos modelos (Paredes et al., 2017). Por este motivo, uno de los retos pendientes en la aplicación de estas técnicas es realizar la fusión entre la modelización y el razonamiento teórico basado en los fundamentos de la microeconomía (Athey & Imbens, 2019).

Desde una perspectiva microeconómica, solamente hemos encontrado la propuesta de Brathwaite et al. (2017) para fusionar este tipo de modelos con la teoría económica. Estos autores sugieren que los modelos de aprendizaje automático pueden capturar los límites cognitivos en la toma de decisiones que otros modelos no incluyen, lo que permite modelizar las decisiones de los individuos entendiendo que estas pueden violar el principio de racionalidad o alejarse de las que maximizan la utilidad del individuo. Este aspecto ha sido demostrado en artículos relevantes de la economía conductual (Tversky y Kahneman, 1989). Este aspecto contrasta con los modelos econométricos de elección discreta porque el modelo logit multinomial (MNL) asume que los individuos toman decisiones racionales y buscan maximizar su utilidad esperada al elegir entre varias alternativas (McFadden, 1973). Brathwaite et al. (2017) explican que el método de los árboles de decisión puede contribuir a la modelización de decisiones basadas en procesos de decisión no compensatorios. En otras palabras, estos modelos se denominan así porque no permiten que los atributos negativos de una alternativa considerada sean compensados por los atributos positivos. Concretamente, Brathwaite et al. (2017) establecen que los árboles de decisión pueden representar este tipo de procesos basados en disyunciones de conjunciones (Hauser et al., 2010). Esto significa que, bajo este tipo de proceso de decisión, una persona considerará cualquier alternativa que cumpla al menos una de un conjunto dado de condiciones conjuntivas. Observan en sus estimaciones que el modelo logra una mejora en la predicción de las decisiones de los individuos en comparación con el modelo logit multinomial.

Este Artículo trata de esclarecer el funcionamiento de este tipo de algoritmos en el contexto del mercado laboral. Concretamente en el ámbito del autoempleo. Esto nos permitirá comprender mejor las decisiones de emprendimiento del individuo, tomando una perspectiva distinta a la establecida por los métodos econométricos que asumen maximización de la utilidad y racionalidad en las decisiones de los agentes. Por ello, la pregunta fundamental que trata de responder este Trabajo es si las técnicas de Machine Learning, más concretamente los árboles de decisión, son más precisos en las predicciones en las decisiones en el mercado laboral que un modelo logit multinomial. La comparación entre ambos modelos sugerirá que tipo de proceso de decisión está detrás del emprendimiento en España. Este artículo está estructurado de la siguiente manera. En la Sección 2, establece una revisión de la literatura sobre las distintas variables que han mostrado ser condicionantes en la elección del autoempleo en España. La Sección 3, presenta los datos empleados y describe las variables incluidas en el análisis. La Sección 4, explica la metodología empleada. La Sección 5 explica los principales resultados obtenidos. Finalmente, la Sección 6 establece las conclusiones fundamentales del Articulo.

Revisión de la literatura

Como demuestra la literatura, el trabajo por cuenta propia constituye una alternativa muy interesante ante las situaciones de precariedad en España (Congregado et al., 2010; Cueto et al., 2015). Por otro lado, es una alternativa a la que otros individuos pueden optar debido a sus actitudes hacia el mercado laboral (Simoes et al., 2016). Por este motivo, es imprescindible seguir estudiando el fenómeno del autoempleo, ya que constituye un elemento muy relevante en el mercado laboral (Parker, 2004; Simoes et al., 2016; Molina, 2020). En este sentido, el estudio del autoempleo se puede aproximar desde una perspectiva agregada (Barrado & Molina, 2015) o individualizada (Molina et al., 2016a), atendiendo a factores externos o internos. Desde la perspectiva microeconómica, encontramos una gran cantidad de literatura que señala los distintos factores sociodemográficos determinantes para la entrada en el autoempleo (Simoes et al., 2016; Molina, 2020). Entre estos, destacan los siguientes factores a nivel individual: género, edad, distribución del tiempo laboral, desplazamientos del hogar al trabajo o procesos de transmisión intergeneracional (Dunn & Holtz-Eakin, 2000; Gimenez-Nadal et al., 2012; Koellinger et al., 2013; Velilla et al., 2018, 2020; Campaña et al., 2020; 2018, Molina, 2020; Belloc et al., 2022; Giménez-Nadal et al., 2020, 2022a, 2022b).

En España, existe una gran cantidad de literatura sobre los factores determinantes del autoempleo. En general, esta literatura ha señalado dos aspectos relevantes: el desempleo y la situación financiera del hogar (Alba-Ramírez, 1994; Carrasco, 1999; Congregado et al., 2010; Cueto et al., 2015; Molina et al., 2016b). Alba-Ramírez (1994) encontró un aumento en la propensión al autoempleo en España cuando la duración del desempleo aumenta. Carrasco (1999) estimó modelos logit multinomial y de riesgos competitivos discretos utilizando datos de una muestra longitudinal de hombres españoles para el período 1985-1991. Los resultados sugieren que el desempleo aumenta la probabilidad de entrar en el autoempleo. Además, recibir beneficios por desempleo reduce significativamente la probabilidad de entrar en el autoempleo. Las restricciones de liquidez también son importantes para determinar la propensión al emprendimiento empresarial, pero solo para aquellos que se convierten en trabajadores autónomos con empleados.

Congregado et al. (2010), en un estudio sobre el autoempleo en España, también documentan que, durante los períodos de recesión, los individuos se ven empujadas a comenzar un negocio como forma de empleo. En general, esta investigación sugiere que, durante los períodos de auge económico, existen pocas oportunidades de trabajo asalariado seguro para los trabajadores por cuenta propia, por lo que muchos permanecen en este tipo de trabajo. Por otro lado, durante las recesiones, muchas personas buscan el autoempleo como forma de empleo debido a la falta de otras opciones de trabajo. Cueto et al. (2015), analizan la relación existente entre el desempleo y el trabajo por cuenta propia en España incorporando la dimensión espacial. Para ello recurren al modelo espacial de Durbin, observan que existe un efecto sobre el incentivo en el autoempleo cuando el desempleo de una región vecina aumenta. Estos resultados, sugieren que en España el autoempleo puede verse como un refugio en coyunturas de mayor desempleo. Molina et al. (2016b), analizan la decisión de convertirse en emprendedor en España con los microdatos españoles de la Encuesta Financiera de las Familias de 2011. Estiman mediante modelos de regresión de máxima verosimilitud binaria de acuerdo el papel que

juegan las finanzas del hogar en esta decisión. Los resultados muestran que los activos del hogar y la seguridad financiera que representan afectan al emprendimiento y a la intención de los individuos para emprender.

Además del efecto de las finanzas familiares y el desempleo, la literatura del autoempleo en España señala otros factores que condicionan la propensión al emprendimiento en España como la reputación (Liñán et al., 2011), la transmisión intergeneracional (Ferrando-Latorre et al., 2019), la formación adquirida (Salas-Velasco, 2023) o el estrés laboral (Gimenez-Nadal & Ortega-Lapiedra, 2010). Gimenez-Nadal & Ortega-Lapiedra (2010) realizan una comparación entre trabajadores asalariados y por cuenta propia en relación con el estrés por falta de tiempo. Para ello, utilizan datos de la Encuesta de Uso del Tiempo en España (2002/03). Con la incorporación de indicadores objetivos de asignación de tiempo, el autoempleo aumenta el estrés por falta de tiempo percibido por los hombres. Los autores concluyen que no solo la cantidad sino también la calidad del tiempo libre es importante en el autoempleo.Liñán et al. (2011), investigan sobre los elementos cognitivos ambientales que pueden explicar las diferencias regionales en las intenciones emprendedoras en España. Partiendo de datos sobre 549 estudiantes de último curso universitario de dos regiones españolas – Andalucía y Cataluña - observan que la valoración del emprendimiento en cada región ayuda a explicar las diferencias regionales en las intenciones empresariales. En este sentido, la valoración social del emprendedor fue mayor en la región más desarrollada (Cataluña), lo que afectó positivamente a las normas subjetivas percibidas y el control conductual. En Andalucía, la influencia de la valoración percibida del emprendedor en el entorno cercano fue más importante, afectando la actitud hacia el comportamiento y las normas subjetivas.

Ferrando-Latorre et al. (2019), documentan la existencia de procesos de transmisión intergeneracional en el trabajo por cuenta propia en España. Estos autores utilizaron datos de la Encuesta de Finanzas de Hogares (Banco de España) para los años 2002, 2005, 2008, 2011 y 2014, para identificar a los empresarios como trabajadores autónomos. La actividad empresarial de los individuos se estudió como función de las características demográficas y laborales individuales y parentales. Observan una correlación significativa entre la actividad empresarial de padres e hijos, que persiste durante la década estudiada, que incluye los años de la recesión económica. Documentan diferencias en la transmisión intergeneracional, siendo esta más fuerte en el caso de los emprendedores que para los empleados. Salas-Velasco (2023), analizan el efecto de la educación universitaria sobre el emprendimiento en España. Los resultados muestran que los graduados en odontología, fisioterapia, arquitectura, derecho, bellas artes, farmacia y psicología son los más propensos a convertirse en autoempleados. Por otro lado, los individuos financieramente alfabetizados (titulados universitarios en economía y finanzas) tienen menos probabilidades de comenzar sus propios negocios.

En resumen, la literatura del autoempleo en España ha identificado varios factores que influyen en la propensión al emprendimiento, entre ellos se destacan el desempleo y la situación financiera en el hogar. Además, otros factores como la reputación, la transmisión intergeneracional, la formación adquirida y el estrés laboral también han sido identificados como importantes. En conclusión, la literatura existente ha documentado que el autoempleo es una opción atractiva para muchas personas en España, especialmente en momentos de crisis económica y alta tasa de desempleo. Además, los

resultados sugieren que la promoción del emprendimiento no puede ser una solución única y universal, sino que debe adaptarse a las particularidades de cada región y a las características individuales de los emprendedores.

Datos

Como se ha demostrado en la literatura sobre el emprendimiento, existe una gran variedad de circunstancias sociodemográficas individuales que pueden condicionar la decisión de emprender (Molina, 2020). Por este motivo, recurrimos a los datos de la base de datos European Union Statistics on Income and Living Conditions (EU-SILC) del año 2019. Esta base de datos, realizada por Eurostat, recopila información sobre individuos y hogares desde el año 2003 en 20 países europeos. En este trabajo únicamente utilizaremos estos datos que recogen los hogares e individuos de España, con el objetivo de profundizar en estudio del autoempleo en este país.

Las variables que incluiremos en este estudio son variables relativas a los encuestados, individuos en edad de trabajar, y variables relativas al hogar de estos. En relación con las variables individuales del encuestado disponemos de variables sociodemográficas relevantes como el sexo, el estado civil, la edad del encuestado y el máximo nivel de formación adquirido. Para captar el género del encuestado, recurrimos a una variable dicotómica que muestra con valor 1 si es varón y con valor 0 si es mujer. Podemos conocer el estado civil del encuestado, representando en una variable dicotómica con valor 1 si está casado, y 0 en caso contrario. La edad del encuestado, esta medida en años y el máximo nivel de formación, está categorizado de acuerdo con la clasificación International Standard Classification of Education (ISCED). Para mostrar el nivel de formación, creamos dos variables dicotómicas para representar si el máximo nivel adquirido es la educación secundaria (tomando valor 1 en ese caso, y 0 en el contrario) o si el individuo cuenta con educación universitaria (tomando valor 1 en ese caso, y 0 en el contrario).

Respecto a las variables sociodemográficas del hogar, conocemos el tamaño familiar, el número de hijos en el hogar[1] y la renta familiar disponible anual en euros (dividida por 1000). Además, es preciso mencionar que hacemos uso del módulo especial de la base de datos, Special Module on Intergenerational Transmission of Disadvantages (ITD), household composition and evolution of income. Gracias a este módulo podemos tener acceso a las características del hogar cuando el encuestado tenía 14 años. De este modo, podemos controlar los aspectos relacionados con los procesos de transmisión intergeneracional que pueden influir en algún modo sobre las decisiones en el mercado laboral en España (Ferrando-Latorre et al., 2019). Concretamente, conocemos la situación laboral de, al menos el padre o de la madre del encuestado cuando este tenía 14 años. Del mismo modo, podemos conocer la situación financiera del hogar percibida por el encuestado cuando este tenía 14 años y el nivel educativo máximo adquirido por los padres cuando el encuestado tenía 14 años. Por otro lado, cabe mencionar que podemos encontrar individuos en nuestra muestra cuyos padres no estuvieron presentes o solamente uno de los progenitores estuvieron presentes cuando los encuestados tenían 14 años.

[1] Establecemos 3 como máximo número de hijos posibles y 5 como tamaño familiar máximo posible.

Para conocer la situación laboral del padre y/o la madre creamos dos variables dicotómicas, una que refleja con valor 1 si el padre era trabajador autónomo (0 en caso negativo) y otra que refleja con valor 1 si la madre era autónoma (0 en caso negativo). Para la situación financiera percibida en el hogar cuando el individuo tenía 14 años, disponemos de una variable categórica que clasifica en 6 categorías la situación percibida: muy mala, mala, moderadamente mala, moderadamente buena, buena y muy buena. Por motivos de simplicidad, agrupamos estas categorías en 1 variables dicotómica. Esta variable representa con valor 1 si el individuo pertenecía a un hogar con una situación financiera percibida muy buena, buena o moderadamente buena, y con nivel 0, si era muy mala, mala o moderadamente mala. Respecto al nivel educativo de los padres, creamos cuatro variables que reflejan para ambos padres su máximo nivel formativo adquirido. Dos variables, para el padre y la madre, que reflejan si el máximo nivel adquirido es educación secundaria (0 en caso negativo), y dos variables, una para el padre y otra para la madre, que reflejan si el máximo nivel es la formación universitaria (0 en caso contrario).

Finalmente, también podemos conocer la situación laboral de los encuestados, lo cual nos permitirá comprobar de manera fiable, la precisión de las técnicas de IA en la clasificación y predicción de la situación laboral de los individuos. Dicha información se encuentra clasificada de acuerdo con las siguientes categorías: 1) Trabajador por cuenta ajena (tiempo completo); 2) Trabajador por cuenta ajena (tiempo parcial); 3) Trabajador autónomo (a tiempo completo, incluye trabajadores en el negocio familiar); 4) Trabajador autónomo (a tiempo parcial, incluye trabajadores en el negocio familiar); 5) Desempleado; 6) Estudiante; 7) Retirados del mercado laboral; 8) Discapacitados para trabajar; 9) Realizando el servicio militar obligatorio o 8 responsabilidades comunitarias; 10) Encargado de realizar las tareas del hogar; 11) Otro tipo de personas inactivas. Para nuestro estudio, únicamente analizaremos aquellos individuos que cumplan alguna de las siguientes condiciones: 1,2,3,4,5. Creamos una variable de respuesta categórica, con tres categorías y que toma 3 valores para cada una de ellas. El valor 1, representa a los individuos que están desempleados. El valor 2 a los asalariados y el valor 3 a los autoempleados. De tal forma, que incluimos en nuestro análisis aquellas personas que no están incapacitadas de manera irrevocable para la participación en el mercado laboral. Para un resumen de las variables empleadas en el análisis y su descripción, ver Tabla 1. Se puede observar en la Tabla 2 el resumen estadístico de la muestra empleada.

Metodología

Como ya se ha mencionado previamente, para realizar nuestro modelo microeconómico, recurrimos a modelos de elección discreta (Einhorn, 1970; Tversky, 1972; McFadden, 1973; Hauser et al., 2010; Brathwaite et al., 2017). En este sentido, debemos plantear un marco teórico sobre el cual optar por una metodología consistente (Brathwaite et al., 2017). Para ello, recurrimos a los fundamentos de la toma de decisiones desde una perspectiva micro, pues el objeto de estudio es la decisión de emprender como autoempleado. En la literatura sobre procesos de elección discreta encontramos distintos conceptos relacionados con los modelos de decisión no compensatorios[2] como la

[2] En el contexto de la elección discreta, donde los individuos deben elegir una opción de un conjunto finito de alternativas, un modelo no compensatorio se refiere a un modelo en el que los atributos de una alternativa no se combinan de manera lineal y ponderada para determinar la utilidad de la alternativa. En

eliminación por aspectos (Tversky, 1972) o las reglas conjuntivas y disyuntivas (Brathwaite et al., 2017). Recurrimos a este tipo de modelos, debido a que son los que funcionan como enlace entre la teoría de la decisión y los algoritmos de IA (Brathwaite et al., 2017). En particular, podemos entender que la modelización por árboles de decisión se basa un modelo de decisión conocido como "disjunctions-of-conjunctions" (Hauser et al., 2010; Brathwaite et al., 2017). Esto es, bajo este tipo de proceso de decisión, una persona considerará cualquier alternativa que cumpla al menos una de un conjunto dado de condiciones conjuntivas. Cada condición puede tener diferentes requisitos que componen la conjunción (Hauser et al., 2010; Brathwaite et al., 2017). En el contexto específico de esta investigación, suponemos que los individuos consideren diferentes combinaciones de factores al evaluar la viabilidad del autoempleo como una alternativa laboral.

Un árbol de decisión es una técnica propia del aprendizaje automático o machine learning (ML), en términos generales[3], se trata de una concatenación de preguntas lógicas ("si, entonces" que ayuda a realizar predicciones (Loh, 2011; Brathwaite et al., 2017). Para el caso de la elección del autoempleo, los árboles de decisión podrían representar las diferentes consideraciones que un individuo toma en cuenta al momento de decidir si emprender o no un negocio propio. Cada "nodo" del árbol representaría una variable o consideración específica, como el nivel de ingresos, la experiencia previa en el sector, la carga de trabajo, la disponibilidad de recursos, entre otras. Cada "rama" del árbol representaría una decisión que el individuo podría tomar en función de la variable correspondiente. Por ejemplo, si el nodo representa el nivel de ingresos, una rama podría indicar que, si el nivel de ingresos proyectado supera cierto umbral, entonces el individuo optaría por el autoempleo. En este sentido, los árboles de decisión se ajustan bien a la teoría de la decisión no compensatoria, ya que permiten modelar procesos de decisión en los que los individuos pueden tener preferencias estrictas sobre determinadas variables y no estar dispuestos a comprometerlas en función de otras. Además, al utilizar los árboles de decisión en el análisis de la elección del autoempleo, se puede obtener información sobre las variables más relevantes que influyen en la decisión de emprender o no, lo cual puede ser útil para diseñar políticas que fomenten el autoempleo.

Por otro lado, y desde una perspectiva convencional, este tipo de análisis se ha llevado a cabo con modelos logit multinomiales basados en la teoría de la utilidad aleatoria, la cual establece que los individuos toman decisiones racionales al elegir entre varias alternativas disponibles (McFadden, 1973). De este modo, en el contexto de la elección del autoempleo, esto significa que los individuos elegirán el autoempleo si perciben que les proporciona una mayor utilidad esperada que el empleo asalariado, y esta percepción dependerá de factores como la rentabilidad esperada del negocio y sus preferencias personales. Esta disyuntiva nos permite comprobar que proceso se adecúa mejor a la decisión del autoempleo, por un lado, estimaremos un modelo logit multinomial, por otro un árbol de decisión bayesiano.

---

lugar de eso, estos modelos asumen que las alternativas son evaluadas de manera no compensatoria, lo que significa que ciertos atributos pueden ser más importantes que otros y pueden llevar a la elección de una alternativa incluso si ésta tiene peores atributos en otros aspectos (Einhorn, 1970).

[3] Es decir, arboles de regresión, arboles de clasificación o listas de clasificación (Brathwaite et al., 2017)

MODELO LOGIT MULTINOMIAL

En este trabajo consideramos la variable relativa al mercado laboral como una variable de respuestas múltiples que no sigue un orden determinado. Por este motivo, el modelo logit multinomial es la metodología más utilizada para modelizar este tipo de decisiones[4]. En estos modelos las diferentes alternativas posibles que puede tomar la variable dependiente discreta es Y, pueden clasificarse como j = 0,1,2,…J. Cada una de estas alternativas tiene asociada una probabilidad determinada. Esta probabilidad asociada a cada categoría viene determinada por una serie de regresores. Esta probabilidad viene determinada por:

$$p_{ij} = pr(\,Y_i = j) = \frac{\exp\,(x'_i\beta_j)}{[1+\sum_{h=1}^{J}\exp\,(x'_i\beta_h)]}, \text{ j = 1,2,…,J}$$

En la expresión anterior, cada alternativa tiene su propio vector de coeficientes. En nuestro modelo, el vector X está compuesto por las variables explicativas de la Tabla 1 y la Y hace referencia a la variable Status de la Tabla1. La estimación se lleva a cabo por máxima verosimilitud:

$Max\ \sum_{i=1}^{N} l_i\,(\beta)$, dónde

$l_i(\beta) = \sum_{j=0}^{J} 1[Y_i = j]\ln p_{ij}$,

Una consideración relevante es la interpretación de los parámetros en este tipo de modelos, el cálculo de los efectos parciales es complejo. El cambio en la probabilidad de que el individuo *i* elija la opción *j* cuando lo hace una variable continua X_k viene dado por la siguiente expresión:

$$\frac{\partial p_{ij}}{\partial x_k} = p_{ij}[\beta_{jk} - \frac{\sum_{h=1}^{J}\beta_{hk}\exp(x'_i\beta_h)}{[1+\sum_{h=1}^{J}\exp\,(x'_i\beta_h)]}],$$

En la expresión anterior el parámetro β_{jk} no indica información sobre la dirección del cambio en la probabilidad, el cual dependerá de otros elementos. Por este motivo, debemos recurrir a otras medidas como los odds-ratio que presentan una expresión más sencilla:

$$\frac{p_{ij}}{p_{0j}} = \exp(x'_i\beta_j), \text{ j = 1,2,…,J}$$

Dónde el cambio en $\frac{p_{ij}}{p_{0j}}$ ante la variable X_k será $\beta_{jk}\exp\,(x'_i\beta_i)\Delta X_k$, que se simplifica con la aplicación de los log-odds ratios:

$$\ln\left(\frac{p_{ij}}{p_{0j}}\right) = x'_i\beta_j,$$

Sin embargo, en este caso nos interesa conocer la precisión del modelo para adaptarse a la realidad económica. Tradicionalmente, a la hora de evaluar la bondad de ajuste de los modelo logit multinomiales se recurre a distintas mediad como el porcentaje de predicciones correctas, a los pseudo-R^2 y a la log-verosimilitud. Una predicción es

[4] Esto es así debido a las dificultades que presentan otro tipo de modelos en relación con su función de Verosimilitud (Woolridge, 2015).

correcta si la máxima probabilidad estimada se ha obtenido para la categoría de situación laboral correspondiente a su valor real. Por el contrario, si para esta misma observación la mayor probabilidad estimada se obtiene para la categoría cuyo valor no corresponde con el valor real la predicción no es correcta. Para calcular dicho porcentaje creamos una variable binaria que toma valor 1 si la máxima probabilidad se asocia al valor real, y 0 si la predicción no es correcta. A través de la media de esta variable obtenemos la proporción de predicciones correctas del modelo logit multinomial. Respecto a los pseudo-R^2, obtenemos los correspondientes a de McFadden, de Maddala y de Cragg y Uhler. Por otro lado, podemos incorporar a nuestro análisis la log-verosimilitud del modelo. Esta medida, hace referencia a la probabilidad de obtener los datos observados dadas las estimaciones de los parámetros del modelo. Definimos la log-verosimilitud de acuerdo con la siguiente expresión:

$$l_i(\beta) = \sum_{j=0}^{J} 1[Y_i = j] \ln p_{ij},$$

En general, se prefiere el modelo que tenga una log-verosimilitud más alta, ya que esto indica una mejor capacidad de ajuste del modelo a los datos observados. Esto significa que el modelo es capaz de explicar de manera más precisa y adecuada la variabilidad de los datos de respuesta observados, en comparación con otros modelos alternativos.

El problema radica en que estamos comparando dos métodos de estimación y predicción distintos. Por un lado – y como hemos comprobado- los modelos logit multinomiales realizan sus estimaciones mediante la maximización de la verosimilitud. Por otro lado, el árbol de decisión es un modelo predictivo que busca dividir el conjunto de datos en subconjuntos homogéneos con respecto a la variable de respuesta y utiliza reglas de decisión simples para predecir la categoría de respuesta para nuevas observaciones. Por este motivo, estos modelos tienen fundamentos teóricos y objetivos diferentes, lo que complica la comparación entre ambos.

En cuanto a estos métodos de comparación de modelos debemos aclarar varias limitaciones y consideraciones. En primer lugar, estamos analizando medidas precisión del modelo, y no medidas de validación estadística del modelo. En otras palabras, la etapa de validación estadística se puede definir como la evaluación de la generalización de un modelo estadístico (Parady et al., 2021). Por otro lado, la precisión predictiva suele cuantificarse como una función de la discrepancia entre los resultados predichos y observados (es decir, el error de predicción) (Parady et al., 2021).

Arboles de decisión

Como algoritmo de Aprendizaje Automático emplearemos árboles de decisión, concretamente árboles de clasificación (Breiman, 1984). Este tipo de métodos comenzaron a desarrollarse en la década de 1970 ((Breiman, 1984), y se postulan como una alternativa atractiva a métodos convencionales estadísticos que se basan en hipótesis de linealidad y normalidad. Estos métodos constituyen la base principal de otros más avanzados y complejos en los que se combinan múltiples árboles para mejorar (Bosques aleatorios, cita).

El concepto fundamental de los árboles de decisión consiste en la partición del espacio predictor (Breiman, 1984). Es decir, se dividen el conjunto de posibles valores de las

variables explicativas en regiones simples de forma que se pueda representar el proceso en forma de árbol. Originariamente, se parte de un nodo inicial que representa la totalidad de la muestra utilizada. A raíz del nodo original, subyacen dos ramas que dividen la muestra en dos subconjuntos, representados por un nuevo nodo. Este proceso se repite un número finito de veces hasta obtener los nodos terminales, empleados para realizar la predicción. Ya construido el árbol, la predicción se realiza en función de la moda, en los casos de clasificación, y de la media en los casos de regresión. Para realizar la división de cada nodo, se selecciona una variable predictora y se realiza una pregunta dicotómica sobre ella. Por ejemplo, un conjunto podría ser dividido en función de un umbral establecido de renta familiar disponible o del estado civil. Esto se conoce como partición recursiva (Breiman, 1984), el objetivo principal de esta técnica es que los nodos terminales sean homogéneos respecto a la variable dependiente. Finalizado el proceso iterativo, el especio predictor queda dividido en regiones dónde la predicción de la respuesta es constante.

En particular, en este Articulo queremos predecir la variable referida a la situación laboral con el objetivo de comprender las decisiones relativas al autoempleo. Por tanto, debemos construir un árbol de clasificación, dónde la variable dependiente puede tomar los valores 1, 2 y 3, que se refieren a las categorías laborales de nuestra muestra (Desempleado, autoempleado y trabajador por cuenta ajena, respectivamente). Por tanto, se procede a construir el árbol comprobando cual es la categoría modal de cada región. Para ello consideramos la muestra de entrenamiento. En el contexto de la modelización de datos, la muestra de entrenamiento se refiere a una parte de los datos que se utilizan para ajustar y entrenar un modelo de aprendizaje automático. Es decir, el modelo utiliza esta muestra de datos para aprender y ajustar sus parámetros. Luego, se utiliza otro conjunto de datos, conocido como conjunto de prueba o validación, para evaluar el rendimiento del modelo.

La proporción de la muestra que se utiliza para el entrenamiento depende de varios factores, como el tamaño de la muestra total, la complejidad del modelo y la cantidad de información que se espera que aporte cada observación. Una práctica común es dividir los datos en una proporción de 70-30 o 80-20, donde el 70% o 80% de los datos se utilizan para el entrenamiento y el resto para la validación del modelo. En este caso, ajustamos la muestra de entrenamiento en un 80% de la muestra total. Para predecir, diremos que una observación pertenecerá a la categoría modal de la región a la que pertenece. El elemento fundamental en la elaboración de los árboles de clasificación es escoger la partición del espacio predictor y para ello, analizaremos medidas de error.

La metodología utilizada en este estudio se basa en la implementación del algoritmo CART (Classification and Regression Trees) mediante la función rpart() del paquete rpart en R. Para ello, especificamos la respuesta (variable dependiente status laboral) y las variables predictoras para incluir todas las posibles variables explicativas posibles en nuestro modelo. Además, se utilizó una lista de parámetros opcionales para la partición en el caso de clasificación (parms), estableciendo el criterio de error en "information" que corresponde al criterio de error referido a la entropía[5]. Para el caso de los árboles de clasificación se emplean tres medidas distintas para reflejar el error en la región. Siendo p_k con $k = 1,2,3$ a la proporción de observaciones (en la muestra de entrenamiento) en

[5] Empleamos esta medida debido a que incrementa la precisión de las clasificaciones del modelo.

la región que pertenecen a la categoría 1, 2 o 3. Podemos obtener en primer lugar, la proporción de errores de clasificación:

$1 - Max_k(\, p_k)$,

El índice de Gini

$\sum_{k=1}^{3} p_k(1 - p_k)$,

O la entropía:

$-\sum_{k=1}^{3} p_k \ln(p_k)$,

En otro orden, controlamos el algoritmo de partición del espacio predictor optimizando el parámetro de complejidad, α. Este parámetro comprendido entre 0 y 1, es una medida de la profundidad del árbol. Es decir, un árbol con un valor de complejidad 1 es un árbol sin divisiones y un árbol con complejidad 0 es un árbol con máxima profundidad. En este caso, implementamos la validación cruzada para seleccionar el valor óptimo del parámetro de complejidad (α) en un árbol de clasificación generado con la función rpart() en R. En primer lugar, se utiliza la función rpart() para crear el árbol de clasificación utilizando una muestra de entrenamiento. Además, se establece el parámetro de complejidad α en 0 para iniciar con un primer valor. Luego, se extraen los valores de cp y los errores de clasificación y su desviación estándar correspondientes a cada nivel de complejidad en la tabla de complejidad del modelo generado. A continuación, se calcula el mínimo error de clasificación más su desviación estándar para cada nivel de complejidad, y se determina el valor óptimo del parámetro de complejidad como el valor de cp que minimiza el valor del error. El resto de los parámetros del modelo quedan determinados por los valores por defecto de la librería de R rpart.

Resultados

En primer lugar, observamos los resultados de la estimación del modelo logit multinomial de la Tabla 2. Este método nos permite identificar las variables que influyen en la elección de cada una de las opciones de empleo: asalariado, autoempleado y desempleado. En el modelo establecemos la categoría 2 (asalariado) como variable de referencia para comparar con los otros estados laborales. En un primer momento, para el caso del autoempleo, resultan estadísticamente significativas las siguientes variables: ser varón, tamaño familiar, edad, estado civil, buena situación financiera en el pasado, formación universitaria de la madre, formación universitaria del padre, renta familiar disponible, padre autoempleado cuando el encuestado tenía 14 años y formación secundaria no obligatoria del padre. Es preciso mencionar, que estas son las variables que resultan significativas respecto a la categoría de referencia 2 (asalariado), es decir, las variables significativas entre la decisión de ser autónomo respecto al de ser empleado por cuenta ajena. Esto implica que estas variables no tienen por qué resultar significativas en el caso de la categoría 1, respectiva al desempleo. Para este caso en particular, observamos las siguientes variables estadísticamente significativas: género, número de hijos, tamaño familiar, edad, educación secundaria no obligatoria, educación universitaria, situación financiera en el pasado, formación universitaria del padre, formación secundaria no obligatoria del padre y de la madre y madre autoempleada.

Para la interpretación de los parámetros podemos observar la Tabla 3 que incluye la estimación del modelo de acuerdo a la ratio de riesgo relativo (RRR). Es una medida utilizada en el modelo de regresión logística multinomial, la cual se obtiene al exponenciar los coeficientes de la regresión logística multinomial de la Tabla 2. El RRR indica cómo cambia el riesgo de que el resultado se encuentre en el grupo de comparación, en comparación con el riesgo de que el resultado se encuentre en el grupo de referencia, a medida que aumenta la variable en cuestión. Un RRR > 1 indica que el riesgo de que el resultado se encuentre en el grupo de comparación, en comparación con el riesgo de que el resultado se encuentre en el grupo de referencia, aumenta a medida que aumenta la variable. En otras palabras, el resultado de comparación es más probable. Por el contrario, un n RRR < 1 indica que el riesgo de que el resultado se encuentre en el grupo de comparación, en comparación con el riesgo de que el resultado se encuentre en el grupo de referencia, disminuye a medida que aumenta la variable. En general, si el RRR < 1, el resultado es más probable que esté en el grupo de referencia. De este modo, observamos que es más probable encontrar en el grupo de autoempleados a hombres, casados, individuos con buena situación financiera en el pasado, individuos cuyos padres (padre y/o madre) tenían educación universitaria en el pasado y aquellos encuestados cuyo padre fue autoempleado en el pasado. De otra manera, la probabilidad de ser autoempleado respecto al grupo de referencia disminuye cuanto mayor es la renta familiar disponible y/o cuando el padre disponía de educación secundaria no obligatoria como máximo nivel de formación. Es posible la cuantificación de las probabilidades de la asociación entre el autoempleo y cada variable explicativa mediante el cálculo de los efectos marginales. Su cálculo nos permite observar que la probabilidad predicha por este modelo de que un individuo sea autoempleado y que su situación financiera en el pasado fuese buena es del 14%, o que el autoempleado tuviese un padre autoempleado cuando este tuviese 14 año, en cuyo caso la probabilidad predicha es del 22,34%.

En cuanto a la capacidad predictiva del modelo logit multinomial observamos que presenta una precisión del 72,65%. Es decir, el modelo asocia la mayor probabilidad a la categoría laboral que el individuo realmente desempeña 72,65 veces de cada 100 predicciones. En términos de nuestra muestra el modelo acierta el estatus laboral de la predicción de 7834 individuos de 10783. No obstante, aunque pueda parecer una precisión elevada de los datos la medida de precisión puede ser un indicador erróneo de la capacidad predictiva del modelo. Si observamos la capacidad de predecir a los individuos autoempleados correctamente, observamos que únicamente acierta en 16 casos de 1493 posibles, es decir, muestra una precisión del 0,14% en la predicción de autoempleo. Estos datos sugieren que la capacidad predictiva de los modelos logit multinomiales es baja, al menos en el caso del autoempleo.

En el árbol de decisión, contemplamos dos nodos terminales que determinan la región de los autoempleados. Siguiendo las reglas que marca el algoritmo del árbol de regresión el perfil que dibuja el árbol de clasificación de los autoempleados es doble. Por un lado, el árbol identifica como autoempleados a aquellos individuos que cumplen con las siguientes condiciones: individuos situados en un nivel de renta familiar disponible entre 8,1 y 16,6 mil euros, pertenecientes a una familia compuesta por dos o más individuos, cuyo padre fue autoempleado y que pertenecían a un hogar cuya situación financiera era buena. Por otro lado, encontramos otro perfil para el autoempleo en España: individuos con una renta familiar disponible menor a 8,1 mil euros, pertenecientes a una familia

compuesta por dos o más miembros, con un nivel educativo máximo básico y descendientes de una madre autoempleada cuando el individuo tenía 14 años. En la Figura 1 se puede observar el resto de las condiciones asociadas a las otras categorías del status laboral. De la elaboración de este árbol de clasificación a partir de la muestra de entrenamiento, realizamos predicciones para el 20% restante de la muestra original. Como se puede apreciar en la Tabla X, obtenemos una precisión en las estimaciones del 73,81%, ligeramente superior a las estimaciones del modelo logit multinomial. Cabe mencionar que las predicciones en este caso se realizan sobre la muestra de prueba, en otras palabras, se consigue una mejor predicción con una menor proporción de datos. La tasa de información nula (No Information Rate) representa la precisión que se lograría si se predijera siempre la clase más común del conjunto de datos, en este caso, es del el 72,23%. En el caso del autoempleo, el algoritmo realiza un 47,62% de predicciones correctas. Es decir, si un individuo cumple con las reglas establecidas antes, clasificaremos correctamente como autoempleado un 47,62% de las veces. Esta cifra, permite observar el atractivo de los árboles de decisión en tareas de clasificación. En predicción de autoempleados el árbol de clasificación es 320 veces más preciso que el modelo logit multinomial.

En la Tabla 4, también observamos las estadísticas por categoría. Estas incluyen la sensibilidad, que es la proporción de casos positivos correctamente clasificados, y la especificidad, que es la proporción de casos negativos correctamente clasificados. También se proporcionan valores predictivos positivos y negativos, que representan la proporción de casos positivos y negativos correctamente clasificados entre todas las predicciones positivas y negativas, respectivamente. En la Tabla 5, observamos las mismas medidas, pero para la predicción realizada por el modelo logit multinomial. Como se puede observar las diferencias son considerables. En todas las medidas, el modelo de aprendizaje automático mejora al modelo logit multinomial. Para el caso del autoempleo, el árbol de decisión acierta el 47,6% de sus predicciones positivas y el 86,4% de sus predicciones negativas, las cifras respectivas para el modelo logit multinomial es de 0,14% y 86,1%. El árbol de decisión clasifica correctamente el 3,3% de los autoempleados de la muestra de prueba, mientras que el logit multinomial, clasifica el 1% de los autoempleados como tal en la totalidad de la muestra. En cuanto a la especificidad, el árbol de clasificación clasifica correctamente el 99,4% de los individuos que no son autónomos, frente al 86,2% en el modelo econométrico.

Conclusión

Este Artículo realiza dos contribuciones relevantes al estudio del mercado laboral en España. En primer lugar, mediante la aplicación de algoritmos de Aprendizaje Automático logramos elaborar perfiles sociodemográficos para el autoempleo en España. Concretamente, observamos que el algoritmo clasifica como autónomos aquellos individuos que presentan estás condiciones: individuos situados en un nivel de renta familiar disponible entre 8,1 y 16,6 mil euros, pertenecientes a una familia compuesta por dos o más individuos, cuyo padre fue autoempleado y que pertenecían a un hogar cuya situación financiera era buena. Por otro lado, encontramos otro perfil para el autoempleo en España: individuos con una renta familiar disponible menor a 8,1 mil euros,

pertenecientes a una familia compuesta por dos o más miembros, con un nivel educativo máximo básico y descendientes de una madre autoempleada cuando el individuo tenía 14 años.

En segundo lugar, documentamos una mayor capacidad de clasificación y predicción del status laboral en los modelos de Aprendizaje automático en comparación con los modelos econométricos de elección discreta tradicionales. En particular, observamos que, en todas las medidas, el modelo de árbol de decisión supera al modelo logit multinomial. Además, el modelo de árbol de decisión tiene una mayor sensibilidad para detectar casos positivos correctamente clasificados, y una mayor especificidad para detectar casos negativos correctamente clasificados. Estos resultados sugieren que la fusión de los fundamentos teóricos de la economía y los algoritmos de inteligencia artificial puede ser una herramienta muy valiosa en la investigación en economía. Además, destaca la importancia de explorar diferentes enfoques y perspectivas en este ámbito para lograr la vinculación del espectro teórico con la modelización por algoritmos.

En conclusión, el modelo de árbol de decisión constituye una alternativa muy atractiva para la clasificación de las observaciones en las diferentes categorías. Este modelo presenta una mayor precisión, una mayor sensibilidad y especificidad, y clasifica correctamente un mayor porcentaje de las observaciones en comparación con el modelo logit multinomial. Este aspecto actúa en favor de los modelos basados en reglas conjuntivas y disyuntivas fundamentados en modelos de decisión no compensatorios. En otras palabras, esto podría sugerir que los individuos no actúan de acuerdo con los principios de racionalidad y maximización de la utilidad esperada, en su lugar, las decisiones podrían verse afectadas por sesgos cognitivos y por la valoración de una serie de atributos en estas decisiones. En resumen, combinar diferentes enfoques y perspectivas en la investigación en economía puede ayudar a mejorar nuestra comprensión del mercado laboral y las decisiones individuales que influyen en él. Esto abre nuevos horizontes y posibilidades para explorar la interacción entre los fundamentos teóricos de la economía y los algoritmos de inteligencia artificial.

Referencias

Alba-Ramirez, A. (1994). Self-employment in the midst of unemployment: the case of Spain and the United States. *Applied Economics*, *26*(3), 189–204.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Barrado, B., & Molina, J. A. (2015). Factores macroeconómicos que estimulan el emprendimiento. Un análisis para los países desarrollados y no desarrollados. *Documento de Trabajo 2015-06.* Facultad de Economía y Empresa. Universidad de Zaragoza

Belloc, I., Molina, J. A., & Velilla, J. (2022). Living in rural areas and self-employment. IZA DP, nº 15059.

Brathwaite, T., Vij, A., & Walker, J. L. (2017). Machine learning meets microeconomics: The case of decision trees and discrete choice. *ArXiv Preprint ArXiv:1711.04826*.

Breiman, L. (1984). Classification and Regression Trees (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Campaña, J.C., Giménez-Nadal, J.I. and Molina, J.A. (2020). Self-employed and employed mothers in Latin American families: are there differences in paid-work, unpaid work and child care? *Journal of Family and Economic Issues*, 41, 52-69. https://doi.org/10.1007/s10834-020-09660-5.

Carrasco, R. (1999). Transitions to and from self-employment in Spain: an empirical analysis. *Oxford Bulletin of Economics and Statistics*, *61*(3), 315–341.

Congregado, E., Golpe, A. A., & Carmona, M. (2010). Is it a good policy to promote self-employment for job creation? Evidence from Spain. *Journal of Policy Modeling*, *32*(6), 828–842. https://doi.org/https://doi.org/10.1016/j.jpolmod.2010.09.001

Cueto, B., Mayor, M., & Suárez, P. (2015). Entrepreneurship and unemployment in Spain: a regional analysis. *Applied Economics Letters*, *22*(15), 1230–1235.

Dunn, T., & Holtz-Eakin, D. (2000). Financial capital, human capital, and the transition to self-employment: Evidence from intergenerational links. *Journal of labor economics*, 18(2), 282-305.

Einhorn, H. J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin*, *73*(3), 221.

Ferrando-Latorre, S., Velilla, J., & Ortega, R. (2019). Intergenerational Transmission of Entrepreneurial Activity in Spanish Families. *Journal of Family and Economic Issues*, *40*(3), 390–407. https://doi.org/10.1007/s10834-019-09613-7

Gimenez-Nadal, J. I., Molina, J. A., & Ortega, R. (2012). Self-employed mothers and the work-family conflict. *Applied Economics*, *44*(17), 2133–2147. https://doi.org/10.1080/00036846.2011.558486

Giménez, J.I. , Molina, J.A. and Velilla, J. (2018). The commuting behavior of workers in the United States: differences between the employed and the self-employed. *Journal of Transport Geography*, 66, 19-29. https://doi.org/10.1016/j.jtrangeo.2017.10.011.

Giménez-Nadal, J.I., Molina, J.A. and Velilla, J. (2020). Commuting and self-employment in Western Europe. *Journal of Transport Geography*, 88. 102856. https://doi.org/10.1016/j.trangeo.2020.102856.

Giménez-Nadal, J. I., Molina, J. A., & Velilla, J. (2022a). Intergenerational correlation of self-employment in Western Europe. *Economic Modelling*, *108*. https://doi.org/10.1016/j.econmod.2021.105741

Giménez-Nadal, J. I., Molina, J. A., & Velilla, J. (2022b). Occupational sorting and the transmission of self-employment between generations. *Applied Economics Letters*, 1-4.

Gimenez-Nadal, J. I., & Ortega-Lapiedra, R. (2010). Self-employment and time stress: The effect of leisure quality. *Applied Economics Letters*, *17*(17), 1735–1738.

Hauser, J. R., Toubia, O., Evgeniou, T., Befurt, R., & Dzyabura, D. (2010). Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, *47*(3), 485–496.

Koellinger, P., Minniti, M., & Schade, C. (2013). Gender Differences in Entrepreneurial Propensity. *Oxford Bulletin of Economics and Statistics*, *75*(2), 213–234. https://doi.org/10.1111/j.1468-0084.2011.00689.x

Liñán, F., Urbano, D., & Guerrero, M. (2011). Regional variations in entrepreneurial cognitions: Start-up intentions of university students in Spain. *Entrepreneurship and Regional Development*, *23*(3–4), 187–215.

Loh, W. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23.

McFadden, D. (1973) Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*. New York: Wiley.

Molina, J. A. (2020). Family and Entrepreneurship: New Empirical and Theoretical Results. In *Journal of Family and Economic Issues* (Vol. 41, Issue 1). Springer. https://doi.org/10.1007/s10834-020-09667-y

Molina, J. A., Velilla, J., & Ortega, R. (2016a). *Entrepreneurial activity in the OECD: Pooled and cross-country evidence*. MPRA Paper 71592.

Molina, J. A., Velilla, J., & Ortega, R. (2016b). The decision to become an entrepreneur in Spain: the role of household finances. *International Journal of Entrepreneurship*, *20*, 57.

Parady, G., Ory, D., & Walker, J. (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. In *Journal of Choice Modelling* (Vol. 38). Elsevier Ltd. https://doi.org/10.1016/j.jocm.2020.100257

Paredes, M., Hemberg, E., O'Reilly, U.-M., & Zegras, C. (2017). Machine learning or discrete choice models for car ownership demand estimation and prediction? *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 780–785.

Parker, S. C. (2004). *The economics of self-employment and entrepreneurship*. Cambridge university press.

Salas-Velasco, M. (2023). Propensity for Self-Employment in a Model of Occupational Choice: Evidence from a Cohort of Recent University Graduates in Spain. *Sustainability*, *15*(4), 3400.

Simoes, N., Crespo, N., & Moreira, S. B. (2016). Individual determinants of self-employment entry: What do we really know?. Journal of economic surveys, 30(4), 783-806.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281.

Tversky, A., & Kahneman, D. (1989). Rational choice and the framing of decisions. *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*, 81–126.

Velilla, J., Molina, J. A., & Ortega, R. (2018). Why older workers become entrepreneurs? International evidence using fuzzy set methods. *Journal of the Economics of Ageing*, *12*, 88–95. https://doi.org/10.1016/j.jeoa.2018.03.004

Velilla, J., Molina, J. A., & Ortega, R. (2020). Entrepreneurship among low-, mid-and high-income workers in South America: a fuzzy-set analysis. IZA DP, nº 13209.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Tabla 1. Descripción de las variables utilizadas en el análisis.

VARIABLES	Descripción
Being male	Variable dicotómica que toma valor 1 cuando el individuo es varón y 0 si es mujer
Age	Edad medida en años
Being married	Variable dicotómica que toma valor 1 cuando el individuo está casado y 0 en caso contrario
Edu: School/compulsory	Variable dicotómica que toma valor 1 cuando la formación máxima del individuo es la formación obligatoria, y 0 en caso contrario
Edu: High-school non-compulsory	Variable dicotómica que toma valor 1 cuando la formación máxima del individuo es la formación secundaria no obligatoria, y 0 en caso contrario
Edu: College	Variable dicotómica que toma valor 1 cuando la formación máxima del individuo es la formación universitaria, y 0 en caso contrario
Family size	Tamaño familiar
Number of kids	Número de hijos
Family disposable income	Renta familiar disponible anual en euros, dividida por mil. La renta familiar disponible total se "equipara" para tener en cuenta el impacto de las diferencias en el tamaño y la composición del hogar.
SM: mother self-employed	Variable dicotómica que toma valor 1 si la madre del individuo era autoempleada cuando tenía 14 años, 0 en caso contrario
SM: father self-employed	Variable dicotómica que toma valor 1 si el padre del individuo era autoempleado cuando este tenía 14 años, 0 en caso contrario
SM: financial situation: Good	Variable dicotómica que toma valor 1 cuando el individuo disponía de una situación financiera familiar muy buena, buena o moderadamente buena, y 0 en caso contrario
SM: mother edu: High-school non-compulsory	Variable dicotómica que toma valor 1 cuando la formación máxima de la madre del individuo es la formación secundaria no obligatoria, y 0 en caso contrario
SM: mother edu: College	Variable dicotómica que toma valor 1 cuando la formación máxima de la madre del individuo es la formación universitaria, y 0 en caso contrario
SM: father edu: High-school non-compulsory	Variable dicotómica que toma valor 1 cuando la formación máxima del padre del individuo es la formación secundaria no obligatoria, y 0 en caso contrario
SM: father edu: College	Variable dicotómica que toma valor 1 cuando la formación máxima del padre del individuo es la formación universitaria, y 0 en caso contrario
Status	Variable con tres categorías de situación laboral: 1 desempleado, 2 asalariado y 3 autoempleado

Nota: La tabla muestra el conjunto de variables que utilizamos en el análisis. Las variables que empiezan por SM, son las referidas al módulo especial y representan el entorno del individuo cuando este tenía 14 años.

Tabla 2. Estadística descriptiva.

VARIABLES	(1) mean	(2) sd
Being male	0.517	0.500
Age	44.790	8.417
Being married	0.654	0.476
Edu: School/compulsory	0.319	0.466
Edu: High-school non-compulsory	0.238	0.426
Edu: College	0.443	0.497
Family size	1.873	0.564
Number of kids	0.110	0.355
Employee	0.729	0.445
Self-employed	0.132	0.338
Unemployed	0.139	0.346
Family disposable income	34.308	22.993
SM: mother self-employed	0.080	0.272
SM: mother employee	0.259	0.438
SM: father self-employed	0.225	0.418
SM: father employee	0.750	0.433
SM: financial situation: Bad	0.230	0.421
SM: financial situation: Good	0.770	0.421
SM: mother edu: School/compulsory	0.832	0.373
SM: mother edu: High-school non-compulsory	0.091	0.288
SM: mother edu: College	0.077	0.266
SM: father edu: School/compulsory	0.769	0.421
SM: father edu: High-school non-compulsory	0.110	0.312
SM: father edu: College	0.121	0.326

Nota: Estadística descriptiva de la muestra utilizada en el análisis. El número total de observaciones es 10783. En la columna 1 se observa la media de las variables y en la columna 2 la desviación típica.

Tabla 3. Estimación del modelo logit multinomial.

VARIABLES	(1) Unemployed	(3) Self-employed
Being male	-0.819***	0.525***
	(0.063)	(0.063)
Number of kids	-0.175*	-0.056
	(0.105)	(0.110)
Family size	0.485***	0.147*
	(0.069)	(0.076)
Age	0.026***	0.027***
	(0.004)	(0.004)
Being married	-0.084	0.546***
	(0.071)	(0.079)
Edu: High-school non-compulsory	-0.556***	-0.033
	(0.078)	(0.082)
Edu: College	-0.941***	-0.100
	(0.083)	(0.080)
SM: financial situation: Good	-0.297***	0.322***
	(0.068)	(0.079)
SM: mother edu: College	-0.204	0.369***
	(0.172)	(0.127)
SM: father edu: College	0.400***	0.366***
	(0.129)	(0.108)
Family disposable income	-0.050***	-0.013***
	(0.002)	(0.002)
SM: father self-employed	0.109	0.951***
	(0.078)	(0.067)
SM: father edu: High-school non-compulsory	-0.212*	-0.229**
	(0.128)	(0.115)
SM: mother edu: High-school non-compulsory	0.246*	0.163
	(0.127)	(0.118)
SM: mother self-employed	-0.328**	0.126
	(0.131)	(0.100)
Constant	-1.214***	-3.996***
	(0.236)	(0.248)
Observations	10,783	10,783

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Tabla 4. Estimaciones del modelo logit multinomial con la estimación por ratio de riesgo relativo

VARIABLES	(1) Unemployed	(3) Self-employed
Being male	0.441***	1.691***
	(0.0278)	(0.106)
Number of kids	0.839*	0.946
	(0.0880)	(0.104)
Family size	1.624***	1.158*
	(0.112)	(0.0882)
Age	1.026***	1.028***
	(0.00405)	(0.00415)
Being married	0.920	1.727***
	(0.0655)	(0.137)
Edu: High-school non-compulsory	0.574***	0.968
	(0.0445)	(0.0793)
Edu: College	0.390***	0.904
	(0.0325)	(0.0725)
SM: financial situation: Good	0.743***	1.380***
	(0.0503)	(0.109)
SM: mother edu: College	0.815	1.446***
	(0.141)	(0.183)
SM: father edu: College	1.491***	1.442***
	(0.192)	(0.156)
Family disposable income	0.951***	0.987***
	(0.00230)	(0.00161)
SM: father self-employed	1.116	2.587***
	(0.0867)	(0.172)
SM: father edu: High-school non-compulsory	0.809*	0.795**
	(0.104)	(0.0911)
SM: mother edu: High-school non-compulsory	1.278*	1.177
	(0.162)	(0.139)
SM: mother self-employed	0.720**	1.134
	(0.0943)	(0.113)
Constant	0.297***	0.0184***
	(0.0702)	(0.00456)
Observations	10,783	10,783

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Tabla 5. Estadística del árbol de decisión por categoría

Estadísticas	Árbol de decisión		
	Categoría: Desempleado	Categoría: Asalariado	Categoría: Autoempleado
Sensibilidad	0.234	0.97	0.033
Especificidad	0.969	0.168	0.994
Valores positivos predichos	0.555	0.752	0.476
Valores negativos predichos	0.887	0.687	0.864
Prevalencia	0.138	0.722	0.139

Nota: La Tabla recoge las estadísticas de predicción del árbol de clasificación. Estas incluyen la sensibilidad, que es la proporción de casos positivos correctamente clasificados, y la especificidad, que es la proporción de casos negativos correctamente clasificados. También se proporcionan valores predictivos positivos y negativos, que representan la proporción de casos positivos y negativos correctamente clasificados entre todas las predicciones positivas y negativas, respectivamente. La prevalencia representa la proporción de casos en la clase respectiva en la referencia.

Tabla 6. Estadística del modelo logit multinomial por categoría

Estadísticas	Modelo logit multinomial		
	Categoría: Desempleado	Categoría: Asalariado	Categoría: Autoempleado
Sensibilidad	0.107	0.989	0.01
Especificidad	0.872	0.962	0.862
Valores positivos predichos	0.015	0.709	0.0014
Valores negativos predichos	0.859	0.279	0.861
Prevalencia	0.138	0.722	0.139

Nota: La Tabla recoge las estadísticas de predicción del árbol de clasificación. Estas incluyen la sensibilidad, que es la proporción de casos positivos correctamente clasificados, y la especificidad, que es la proporción de casos negativos correctamente clasificados. También se proporcionan valores predictivos positivos y negativos, que representan la proporción de casos positivos y negativos correctamente clasificados entre todas las predicciones positivas y negativas, respectivamente.