LA CREACIÓN DE EMPRESAS POR INMINGRANTES EN ESPAÑA: UN MODELO DE CLASIFICACIÓN

BUSINESS CREATION BY IMMINGRANTES IN SPAIN: A CLASSIFICATION MODEL

Virginia Navajas Romero
Departamento de Economía. Universidad Loyola Andalucía
vnavajas@uloyola.es
Mª del Carmen López Martín
Departamento de Economía. Universidad Loyola Andalucía
mclopez@uloyola.es
Alfonso Carlos Martínez Estudillo
Departamento de Métodos Cuantitativos. Universidad Loyola Andalucía
acme@uloyola.es

RESUMEN

La población inmigrante a la hora de emprender por cuenta propia presenta características diferentes a las que detecta el emprendimiento nacional debido a las limitaciones legales existentes y a los problemas a los que se enfrentan en su integración en un país diferente al de origen. Este trabajo proporciona evidencia empírica respecto a las particularidades que presenta este tipo de empresas, clasificándolas según la duración de las mismas. La principal novedad de este estudio radica en que se realiza a partir de los datos proporcionados por la Muestra Continua de Vidas Laborales (MCVL), una fuente de carácter administrativo y universal con información de más de un millón de trabajadores y pensionistas.

PALABRAS CLAVE: inmigración, emprendimiento, MCVL, minería de datos, árboles de decisión

ABSTRACT

The immigrant population in undertaking self has different characteristics to that detects the national undertaking due to legal constraints and problems facing their integration in a different country from the original. This paper provides empirical evidence with regard to the particularities that presents this kind of companies, classifying them according to their life time. The main novelty of this study lies in that is based on the data provided by the Continuous Sample of Working Lives (MCVL), an administrative and universal source of more than a million workers and pensioners in Spain.

KEY WORDS: immigration, entrepreneurship, MCVL, data mining, decision tree.

1. INTRODUCCIÓN

En España, la actividad emprendedora y la creación de nuevas iniciativas empresariales, son aspectos esenciales para la futura prosperidad de su economía. El emprendimiento es una vía a través de la cual se puede estimular y conducir la transformación del tejido económico español hacia el nuevo modelo económico donde dominará la competitividad económica (OCDE 2012). Por otro lado, y de forma paralela a esta última reflexión, debe tenerse en cuenta que el colectivo inmigrante emprendedor constituye un grupo especial dentro del colectivo de emprendedores, puesto que, además de otros aspectos, se ve afectado por el diseño de las políticas de extranjería y emprendimiento. Sin embargo, pese a ser un colectivo que puede ser clave para la futura prosperidad española y que puede estimular la transformación del tejido empresarial español, no ha sido objeto de suficiente atención.

Principalmente dos son las cuestiones básicas que obstaculizan el análisis de este colectivo desde el punto de vista metodológico. La primera se debe a que el concepto “inmigrante”, es un vocablo de uso cada vez más habitual, pero sin una definición admitida y utilizada de forma general por todos los países; en consecuencia, desde el punto de vista metodológico, a nivel internacional, no se halla similitud en los censos de población inmigrante ya que no existe heterogeneidad en la presentación de los datos. La segunda cuestión se refiere a la complejidad de fuentes existentes en España, lo que supone una traba significativa para conocer esta realidad.

Desde el punto de vista de las fuentes estadísticas para el estudio de las migraciones y por lo tanto de los inmigrantes emprendedores en España, se puede distinguir,. por un lado, las fuentes que son elaboradas con fines específicamente estadísticos con un objetivo investigador y por otro, las fuentes administrativas, de las que se extraen estadísticas y cuyo fin es la propia gestión administrativa. También se puede distinguir entre fuentes universales y muestrales. Las primeras registran o cuentan todos los movimientos o todos los migrantes, ya sea mediante estadísticas explícitamente diseñadas para ello (censos) o registros administrativos (EVR); las segundas, investigan mediante una encuesta (por ejemplo, la Encuesta de Población Activa) y a partir de ella se deduce el volumen total de migraciones y migrantes en el conjunto de la población.

En nuestro caso concreto, para la realidad que interesa estudiar (los trabajadores inmigrantes autónomos) la fuente de datos que más se ajusta a la población objetivo es la Muestra Continua de Vidas Laborales (MCVL), debido a que es un colectivo muy concreto dentro de los inmigrantes y no existen fuentes de datos que favorezcan su explotación a los niveles de desagregación que se plantean en este trabajo; además, esta fuente permite conocer el perfil socioeconómico de los trabajadores autónomos inmigrantes en el año 2011.

El empleo de esta fuente de datos constituye un trabajo pionero, ya que no existen trabajos que analicen el emprendimiento inmigrante por cuenta propia a través de esta muestra administrativa, que plantea como principal dificultad de manejo la extracción de datos; sin embargo, como a la vez también aporta una muestra muy amplia, permite obtener información del emprendimiento en diferentes fases. En cuanto a la metodología empleada, se propone la utilización de un árbol de decisión para la clasificación de las empresas creadas por este colectivo según su antigüedad. El uso de árboles de decisión frente a otras metodologías se justifica fundamentalmente por el objetivo que se persigue, priorizar la interpretabilidad del modelo frente a la precisión, ya que se desea analizar un fenómeno social.

A partir de las ideas precedentes, el desarrollo de este trabajo sigue la siguiente estructura: en primer lugar se analizan las características de la Muestra Continua de Vidas Laborales en el año 2011 para después presentar la metodología aplicada para la extracción y transformación de los datos de esta base de datos. A continuación, se presenta un modelo basado en árboles de decisión para caracterizar las categorías de empresa. Por último, se exponen las principales conclusiones extraídas del trabajo realizado.

2. METODOLOGÍA

La Muestra Continua de Vidas Laborales (en adelante MCVL) es un conjunto organizado de microdatos anónimos extraídos de registros administrativos que utiliza tres fuentes: los registros administrativos de la Seguridad Social, el Padrón Municipal Continuo y la Agencia Tributaria. Los datos están referidos a aproximadamente un millón de personas, concretamente en el año de referencia que se ha considerado (2011) ascendía a 1.202.387, constituyendo una muestra representativa de todas las personas que tuvieron relación con la Seguridad Social en este año en concreto. En la muestra aparecen personas

con amplio abanico de actividades relacionadas con las fuentes anteriormente mencionadas, tales como aquellas que han cotizado a la Seguridad Social, o las que han cobrado una prestación en el año de referencia debiéndose dicha cotización tanto a la situación de trabajo como a la de percepción de prestaciones por desempleo, bajas por enfermedad, etc.

La MCVL es la primera y única fuente de información construida íntegramente a partir de un registro de carácter administrativo, cubriéndose de este modo el déficit histórico que presenta la explotación estadística en España de este tipo de registros (Cebrián y Toharia, 2007), algo que cuenta con una larga tradición en otros países de nuestro entorno europeo y, especialmente, en el caso de los nórdicos (Tonder, 2008).

Respecto al ámbito temporal de la población de la MCVL, son todos los individuos que han estado en relación con la Seguridad Social en algún momento del año 2011, por lo tanto es de mayor tamaño y tiene una composición algo diferente respecto a lo que se obtendría si se utiliza una fecha fija, que es el criterio habitual en los estudios de migraciones cuyo fuente de datos es la Encuesta de Población Activa (EPA).

Las diferencias entre ambas fuentes de datos se encuentran en tres ámbitos: población incluida, método de obtención de los datos y período al que se refieren:

En cuanto a la población incorporada, la definida para la MCVL no se corresponde conceptualmente con la población activa, lo que presenta la primera diferencia respecto a la EPA ya que en la MCVL las personas se parece más “perceptores de ingresos” que los de la EPA, porque está constituida por individuos que tienen ingresos a su nombre emanados de un empleo presente o pasado, aunque el mismo sea de carácter esporádico, o el empleado haya sido un familiar ya fallecido. Por lo tanto, la población de referencia de la MCVL no se limita únicamente a los que están en alta laboral, sino que incluye también a otras personas que cotizan para pensiones aunque no estén trabajando. Por ejemplo, los que tienen suscrito un Convenio Especial para este fin, los perceptores de prestaciones de desempleo contributivas y no contributivas, aunque no todos ellos acumulan derecho a pensiones. En cambio no están incluidos los demandantes de empleo cuando no reciben estas prestaciones, aunque estén registrados en un servicio público de empleo. Por otra parte, otra cuestión a destacar es que de la MCVL se excluyen aquellas

personas que tienen un sistema de previsión social distinto a la Seguridad Social (en particular los funcionarios de Clases Pasivas) o los que no tienen ninguno (empleo sumergido y algunas actividades marginales) sin embargo éstos sí estarían recogidas en la Encuesta de Población Activa.

En segundo lugar, existen análisis transversales que permiten establecer asociaciones entre variables; esto posibilita realizar estudios de carácter longitudinal y por lo tanto, poder aplicar técnicas de análisis de historia de acontecimientos basadas en el concepto clave de 'curso de vida'. Desde esta perspectiva, la trayectoria de los individuos es interpretada como una serie de transiciones que, a su vez, son el resultado de la situación del individuo en un momento determinado pero, también, de las decisiones tomadas en el pasado y que, en cierta manera, condicionaron transiciones posteriores (Mortimer y Shanahan, 2003). Con la MCVL se pueden reconstruir las trayectorias laborales y, por lo tanto se pueden investigar los mecanismos causales que generan los procesos de cambio en la relación de los individuos con el mercado de trabajo, teniendo en que cuenta que la muestra ofrece las fechas exactas de inicio y fin de las relaciones laborales, posibilitando la aplicación de técnicas de análisis en tiempo continuo. Sin embargo, en el caso de la EPA, si bien también tiene una estructura de panel, se trata de una investigación continua y de periodicidad trimestral dirigida a las familias cuya finalidad principal es obtener datos de la fuerza de trabajo y de sus diversas categorías (ocupados, parados), así como de la población ajena al mercado laboral (inactivos). La muestra inicial es de 65.000 familias al trimestre, quedando reducida en la práctica a aproximadamente 60.000 familias que proporcionan información mediante la realización de entrevistas en las que responden a una encuesta previamente diseñada de acuerdo con los objetivos mencionados.

Finalmente, el periodo de análisis de la MCVL es más amplio que el de la EPA, el cual es limitado en el tiempo ya que se circunscribe a un año y medio de duración (pues es el período que las familias permanecen en el panel de entrevistados), mientras que en la MCVL es como mínimo de cinco años, los cuales irán aumentando con las sucesivas olas.

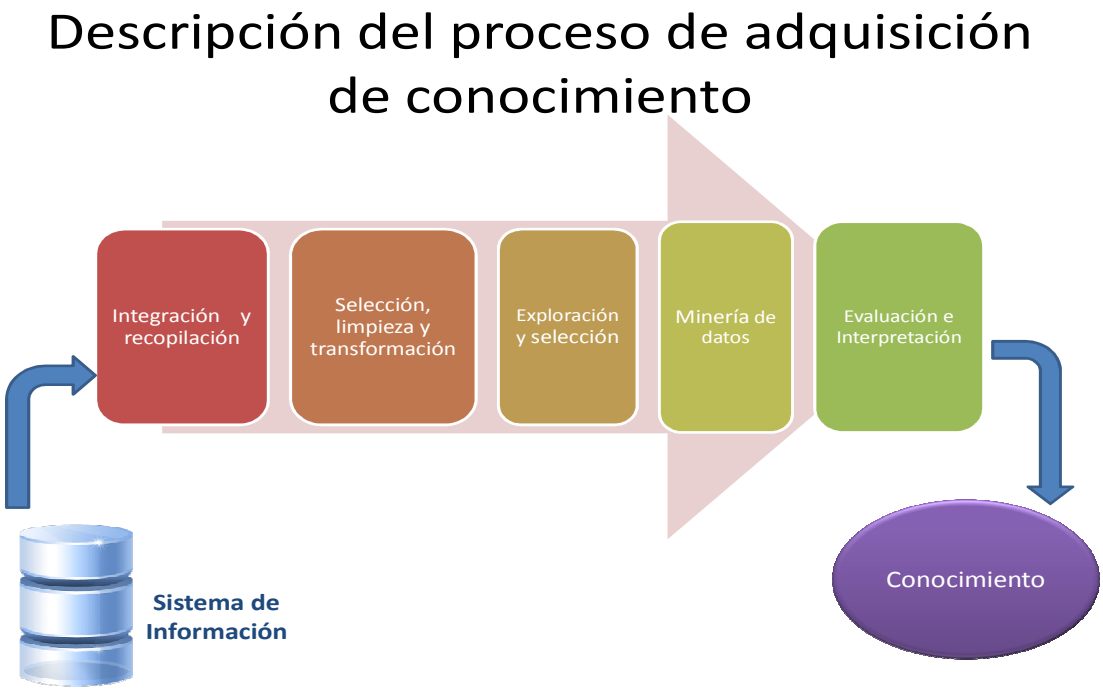
3. EXTRACCIÓN DE LA INFORMACIÓN ACERCA DEL EMPRENDIMIENTO INMIGRANTE A PARTIR DE LA MCVL

El concepto de adquisición de conocimiento en bases de datos (proceso KDD) se define como el proceso donde se extrae conocimiento útil y comprensible de grandes cantidades de datos. La tarea fundamental del descubrimiento de conocimiento en bases de datos es encontrar modelos inteligibles para los mismos, por lo que consiste en la aplicación de métodos para el procesamiento y análisis de datos. Con este objetivo, el proceso se apoya se basa en dos conceptos: “seleccionar y explotar”. Se realiza el tratamiento de un gran volumen de datos mediante diferentes procesos para permitir el análisis de la información extrayendo, por ejemplo, elementos de utilidad o comportamientos interesantes tales como cambios, anomalías, estructuras significativas y patrones de comportamiento, lo cuales podrán luego ser aplicados a nuevos conjuntos de datos.

El proceso de extracción de datos se considera como un proceso reiterado e interactivo: reiterado, ya que en cualquier fase puede plantearse la opción de volver a fases anteriores y donde frecuentemente son necesarias varias iteraciones para extraer conocimiento de alta calidad; e interactivo, porque el usuario debe ayudar en la preparación de datos, en la validación del conocimiento extraído y en todo lo que implica el proceso de minería de datos.

Este proceso se compone de varias fases y, aunque existen distintas maneras de clasificarlas, todas apuntan hacia conceptos similares: fase de integración y recopilación, fase de selección, limpieza y transformación, fase de minería de datos, fase de evaluación e interpretación del modelo, y por último fase de explotación y/o conocimiento.

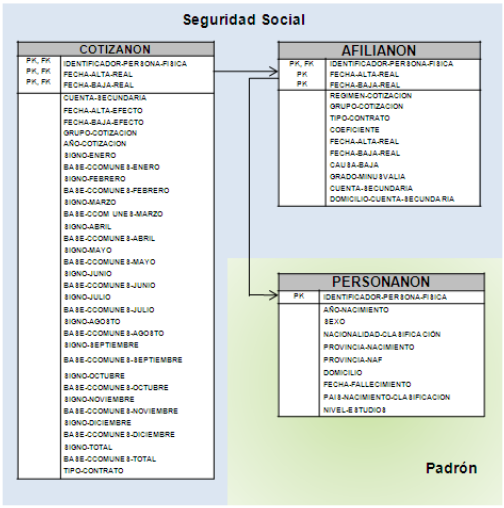
Ilustración 1. Fases del proceso de adquisición de conocimiento



Fuente: Elaboración propia

La versión que se ha utilizado de la MCVL 2011 es la que incluye los datos fiscales. Esta base de datos tiene seis tablas, denominadas: “Afilianon”, “Cotizanon”, “Personanon”, “Convivi”, “Datos fiscales”, y “Preanon”. Cada una de estas tablas está formada por varios ficheros con idéntica estructura entre sí. De todas ellas, para obtener las características del colectivo que interesa, solamente se han utilizado tres tablas: “Afilianon”, “Cotizanon” y “Personanon”. “Afilianon” contiene la información relativa a la actividad realizada por las personas (es la información facilitada por la Seguridad Social); “Cotizanon” incorpora las bases de cotización en el régimen de la Seguridad Social al que pertenece el individuo (en nuestro caso, se elige únicamente los individuos que estaban acogidos al régimen de autónomos); finalmente, “Personanon” recoge la información procedente del padrón municipal.

Ilustración 2. Diagrama de la base de datos de la MCVL con las tablas utilizadas



Fuente: Elaboración propia

3.1. SELECCIÓN, LIMPIEZA Y TRANSFORMACIÓN: TAMAÑO MUESTRAL, FACTORES DE ELEVACIÓN Y TIPO DE MUESTREO

Las personas que forman parte de la Muestra de cada año son cuatro de cada cien de las personas que forman parte de la población de referencia. Esta proporción supuso en 2011 la selección de una muestra de 1,2 millones de personas. Como consecuencia, el factor de elevación es 25.

Sin embargo, debe recordarse, tal como se explicó previamente, que la Muestra es representativa sólo de las personas que se relacionaron con la Seguridad Social en el 2011. Es decir, aunque la MCVL recoge los datos relativos la "historia de la vida laboral" de esas personas en los diez años anteriores, no recoge la información relativa a aquellas personas que en 2011, por ejemplo, han fallecido o han abandonado la vida activa sin causar pensión.

El muestreo es aleatorio simple, sin ningún tipo de estratificación. Se seleccionan todas aquellas personas de la población de referencia de ese año cuyo código de identificación personal contenga, en una determinada posición, unas cifras que en su momento fueron seleccionadas aleatoriamente. Dichas cifras son las mismas todos los años (previamente se comprobó que las cifras situadas en las posiciones que se utilizan en la selección están distribuidas de manera uniforme). Este tipo de muestreo ya ha sido utilizado en otros países, y además, asegura que siempre serán seleccionadas las mismas personas, en la medida que continúen manteniendo relación con la Seguridad Social y no cambien de código de identificación. Por otro lado, proporciona un procedimiento automático para seleccionar las nuevas incorporaciones y

garantiza además que éstas son representativas de las altas en la población. Para comprobación, cada año se realizan una serie de contrastes estadísticos para asegurar que las características básicas de la muestra (sexo, edad, comunidad autónoma de residencia y nacionalidad) se distribuyen de la misma forma que las de la población.

La tabla 1 refleja la evolución de los parámetros tamaño de la población de referencia y tamaño de la muestra para los años 2010 y 2011. De esta forma, la primera ola, correspondiente al año 2010, está compuesta por 1.206.663 individuos, cifra que va disminuyendo en la siguiente hasta llegar a 1.202.387 en el año 2011. Este decremento se debe, lógicamente, al descenso de la población de referencia, que ha pasado de los 30.166.575 a 30.059.675 personas como consecuencia de la disminución en el número de cotizantes y perceptores de prestaciones.

Tabla 1. Tamaños de la población de referencia y muestral en la MCVL

Olas	Población de referencia	Tamaño muestral España
2010	1.206.663	30.166.575
2011	1.202.387	30.059.675

Fuente: Elaboración propia basada en MCVL 2011

3.2. VENTAJAS Y DESVENTAJAS DE LA MCVL PARA EL ESTUDIO DE LOS INMIGRANTES

A pesar de la naturaleza administrativa de los datos de la MCVL, la información considerada ofrece unas posibilidades de análisis científico indudable. No obstante, no se puede olvidar el carácter registral de los microdatos y las limitaciones derivadas del tipo de “medición” que supone para el estudio y la investigación. En este apartado se pretenden señalar las principales ventajas y desventajas de los datos de la MCVL que se deben tener en cuenta a la hora de explotar los datos en la investigación (MTAS, 2006; MTIN, 2009a y 2009b; Lapuerta, 2010).

A partir de la información recogida en la muestra se abre un amplio abanico de áreas de análisis para el estudio. Se destacan a continuación las posibilidades de la MCVL al mismo tiempo que se destacan algunos rasgos de la naturaleza

de los datos.[1]

- El gran tamaño de la muestra, permite niveles de desagregación muy detallados, muy difíciles de estudiar a través de otras fuentes, ya sea por territorio o por perfiles de colectivos específicos.
- La información se amplía y completa con datos padronales y fiscales.
- Al recoger a individuos que han estado en relación con la Seguridad Social en algún momento del año, el grado de cobertura es mayor y tiene una composición algo diferente respecto a lo que se obtendría a una fecha fija, facilitando la presencia de relaciones laborales estacionales o de personas que trabajan ocasionalmente, en particular, facilita la presencia de mujeres y jóvenes que habitualmente quedan fuera de los estudios al permanecer poco tiempo en el mercado de trabajo.

A pesar de lo anterior, existen algunos inconvenientes detectados en la MCVL. De entre éstos, se exponen en este epígrafe los relativos tanto a la extracción de datos como a los relativos a la propia MCVL. La subsanación de los relativos a la extracción de datos se comenta en el apartado 3.3, mientras que las restantes dificultades inherentes a la propia MCVL se exponen a continuación:

- La información se construye bajo criterios conceptuales administrativos, es decir que está diseñada y gestionada con un objetivo diferente al investigador.
- La cantidad y la complejidad de los archivos y de sus datos ralentiza y dificulta los análisis.
- La existencia de valores perdidos, de registros repetidos, valores incoherentes,… en general problemas de depuración, dificultan el tratamiento de la información al introducir elementos agregados en el tratamiento de la información.
- Cada persona puede pertenecer a varias categorías distintas, simultánea o sucesivamente, así por ejemplo se puede estar en situación de cotizante y cobrando un prestación en un mismo año, o estar pluriempleado.

[1] En la bibliografía se recogen algunas de las principales referencias de publicaciones vinculadas al tratamiento y análisis de los datos de la MCVL. Algunas de ellas se citan expresamente en el texto. Las otras se incluyen como referencias temáticas de análisis que pueden ser de interés para el lector.

- El conjunto de individuos que han estado en relación con la Seguridad Social en algún momento del año es de mayor tamaño y tiene una composición algo diferente respecto a lo que se obtendría a una fecha fija.
- Los extranjeros nacionalizados figuran en los archivos de la Seguridad Social como nacionales, por lo que los datos infra estiman la población inmigrante (Argimón y González, 2006: 44), para solucionar esta problemática, se estudia el país de origen de la persona en lugar de la nacionalidad, porque el país de origen queda delimitado en la muestra, mientras que si se emplea la nacionalidad, se perderían datos pues con la obtención de ésta, el inmigrante deja de serlo.
- Las diferencias temporales entre la extracción de la muestra y la información padronal generan disonancias. La variable de nivel educativo debe utilizarse con cautela porque no se actualiza constantemente como las demás. La misma situación se produce con el municipio de residencia, porque puede sufrir modificaciones que determinan la no coincidencia con respecto la situación laboral de un momento dado, por ejemplo, se puede residir en un sitio y trabajar en otro, o haberlo hecho anteriormente.
- La información del empleador concuerda con el momento de extracción de la muestra, y podría, en algunos casos, no ajustarse con la existente en el momento de la relación laboral (MTIN, 2009b).

3.3. EXTRACCIÓN DE LA INFORMACIÓN DE LOS AUTÓNOMOS INMIGRANTES A PARTIR DE LA MCVL

Como se ha indicado, la MCVL incluye cada año algo más de un millón de personas, cada una de las cuales ha tenido de promedio más de una docena de relaciones laborales o situaciones de afiliación diferentes a lo largo de su vida. Hay variables que sólo aparecen una vez por individuo: son las relativas a la persona (nacionalidad, domicilio), pero las relativas al puesto de trabajo y al empleador han de figurar tantas veces como relaciones tenga el individuo. A su vez, las bases de cotización por contingencias comunes están disponibles para cada mes desde 1980; presentadas de manera que se identifique a qué fecha y

a qué relación laboral corresponden[2].

La MCVL está formada por una serie de situaciones que pueden ser clasificadas o delimitadas de diferentes maneras. A efecto de la MCVL cada "relación" es lo que transcurre entre un alta y una baja en Seguridad Social y cada una de ellas da lugar en la Muestra a un registro en el fichero de relaciones o situaciones laborales. En nuestro caso, las "relaciones" que aparecen corresponden a episodios de trabajo por cuenta propia (altas laborales), pero aunque en un principio también incluyen etapas en las que la persona ha estado en alguna de las situaciones que dan lugar a inclusión en la población de referencia: cotizar mediante Convenio Especial, percibir prestación o subsidio de desempleo, etc., en nuestra muestra han sido excluidas.

En el fichero de datos fiscales hay un registro por cada persona y entidad pagadora de la que se han obtenido retribuciones en el año de referencia. Cada registro corresponde a una sola relación, excepto si la misma persona ha tenido varias con diferentes pagadores a lo largo de su vida laboral. Para conectar la información, se localiza como año objetivo 2011; cuando una persona haya trabajado para varios empleadores durante el año, se utiliza como variable clave, la fecha de alta y de baja de la empresa, como se explica posteriormente.

Para distinguir las relaciones que implican una ocupación diferenciada de otras situaciones de afiliación se procedió de la siguiente forma:

Primero, en relación con la información facilitada por la Seguridad Social se realizó un doble filtro: por un lado se trató de diferenciar los diferentes regímenes de cotización (régimen general, hogar, minería, agrario, especiales…) que se encuentran entre los cotizantes en la Seguridad Social y se seleccionó el régimen de cotización de autónomos (que tiene el código el 421). Por otro lado, también se filtró la última fecha de baja, porque en caso de que una persona tuviera varias actividades de autónomos se elige aquella que tenía la fecha de baja real más tardía[3]. De este análisis previo se obtienen

[2] Para poder trabajar de una forma más efectiva, los ficheros de la MCVL, en formato texto, se pasaron a una base de datos en forma de tablas, dentro de un servidor de bases de datos relacional, Microsoft SQL-Server®, estableciendo las claves primarias, foráneas y relaciones entre ellas. El lenguaje de consulta utilizado ha sido TRANSACT-SQL.

[3] Hay que destacar que aquellas empresas que al final de 2011 seguían en actividad han sido recodificadas, adjudicándoles una fecha de baja (en concreto, se ha tomado como fecha de baja 31/12/2011). De esta forma, se pueden calcular los días de vida de las empresas y realizar comparaciones entre ellas.

142.641 registros.

Debido a que existen individuos con diversos registros con la misma fecha de baja, y debido a que cotizan por diferentes actividades se escoge la situación que corresponde al último registro, es decir la fecha de baja real máxima de esa persona (que corresponde a la última actividad emprendedora realizada) primando el criterio de actualidad frente a otros criterios válidos, como por ejemplo la duración en el tiempo de la empresa creada. De esta forma se obtienen 138.676 personas[4].

Para incorporar la información facilitada por la Seguridad Social relativa a las bases de cotización, se cruzaron los datos anteriores con los que ésta facilita empleando para ello la coincidencia en el identificador de persona física[5], en la fecha de baja real y en la fecha de alta real, lo que llevó a un número de 138.663 registros[6].

Por último se cruzó nuestra muestra con la información facilitada por el padrón, mediante el identificador de persona física, y se encontraron 266 personas con el mismo identificador en algunos casos con 4 registros totalmente diferentes, por lo que fueron eliminados de la muestra. De esta forma, el número de personas trabajadoras por cuenta propia finalmente seleccionadas asciende a 138.619.

El dato anterior corresponde con situaciones de trabajo por cuenta propia de población nacional e inmigrante. La MCVL, en función del método en que se ha abordado y acotado la población por cuenta propia, recoge información de 138.619 personas emprendedoras que han tenido una vinculación con la Seguridad Social en 2011 en España. Se trata de un 11,52 %[7] de la población de referencia, mientras que los inmigrantes, son considerados así, de acuerdo

[4] En este proceso cabe considerar situaciones de pluriempleo además de errores en los registros administrativos. En todo caso se ha optado por considerar una única situación actual ya que estos casos afectan a pocas situaciones.

[5] El identificador utilizado está basado en el Número de Identificación Fiscal (NIF), aunque se presenta sometido a transformación para conservar el anonimato de la persona. Los problemas que produce la utilización del NIF, es que hay personas que tienen relación con la Seguridad Social pero carecen de NIF, por ejemplo, menores con pensión de orfandad y algunos extranjeros. En este caso hay que utilizar otro código, pero cuando esa persona adquiere un NIF, el número de identificación personal cambia, lo que pone en peligro su continuidad en sucesivas muestras. Por ello hay que advertir que una parte de las incorporaciones o desapariciones entre una muestra y la del año siguiente serán debidas a cambio en el número de identificación personal, no a fallecimiento o variación en la actividad.

[6] Existían 13 registros sin su correspondiente cotización debido a fallos de recogida de datos por parte de la institución, estos registros fueron eliminados. Además existían dos personas con dos registros duplicados, por lo que ambas fueron eliminadas de la muestra, quedándose en 138.659 personas.

[7] Cifra obtenida mediante la relación de la muestra es 138.619 personas y los datos referentes

con su país de origen buscando evitar evitando la perdida de datos originada por la nacionalidad, suponen el otro 1,11%.

Ilustración 3. Selección de datos

Fuente: Elaboración propia basada en MCVL 2011

3.4. CARACTERIZACIÓN DE LA CATEGORIA DE EMPRESA

En este epígrafe, se presenta un modelo basado en árboles de decisión con el objetivo de caracterizar las categorías propuestas en el modelo GEM[8]. El informe GEM es financiado por el Ministerio de Industria, Energía y Turismo. Uno de los aspectos más importantes que presenta este estudio es el análisis multicrerio basado en las categorías de empresa creadas por inmigrantes en el año 2011 en España, distinguiendo entre 4 clases de empresas: (Naciente, Novel, Junior y Consolidada).

- *Nacientes <8 meses*
- *Noveles (8-18] meses*
- *Junior (18-42] meses*

al padrón que indica el número de personas y no registros (1.202.387).

[8] El observatorio global GEM de la actividad emprendedora, nace para cubrir la falta de información en torno al fenómeno emprendedor, en un momento en que los gobiernos de los países más desarrollados comienzan a vislumbrar la figura del emprendedor como clave para paliar los efectos que se van produciendo en los mercados laborales como consecuencia del propio desarrollo.

o *Consolidadas >42 meses.*

El análisis que se realiza en este artículo se basa en su sistema de clasificación de empresas, se profundiza desde la perspectiva del emprendimiento inmigrante, con el objetivo de conocer cuáles son las características que presentan los inmigrantes emprendedores en cada una de estas categorías, abordándose, por tanto, un problema de clasificación. En la tabla 2. se presenta la distribución de patrones que presentan las cuatro clases.

Tabla 2. Distribución de patrones por categorías

DISTRIBUCIÓN DE PATRONES EN LAS CATEGORÍAS		
CATEGORÍAS	EMPRENDEDORES INMIGRANTES	REGISTROS
NACIENTES	79.200	3.168
NOVEL	48.450	1.938
JUNIOR	56.225	2.249
CONSOLIDACIÓN	140.150	5.606
TOTAL	***324.025***	***12.961***

Fuente: Elaboración propia

3.4.1. DESCRIPCIÓN DE LA VISTA MINABLE

Según se ha indicado en el apartado 3.3 la vista minable está constituida por los registros referentes a los inmigrantes, los cuales corresponden a 12.961 casos. A continuación se detallan las variables independientes que se han tenido en cuenta para el diseño del modelo.

- **Variables independientes**

Luego de la búsqueda bibliográfica y la base de datos inicial, se establece que las variables independientes son:

- *Edad.* Variable de carácter cuantitativo, discreta y numérica. A partir de las variables originales con la información del año, la edad se determina como el número de años entre la fecha de nacimiento y la fecha de referencia de extracción de la muestra MCVL 2011.
- *Sexo*. Es una variable cualitativa, nominal, categórica. Variable original de la MCVL, facilitada por el padrón.

- *Origen geográfico*. Es una variable cualitativa, nominal, categórica (28 categorías). Se obtiene tomando la variable original de país de nacimiento, facilitada por el padrón.
- *Provincia.* Es una variable cualitativa, nominal, categórica (55 categorías). Considerando la provincia donde se desarrolla la actividad de cada individuo dada por el Padrón Continuo, se generan las variables que agregan a los individuos por Comunidad Autónoma, analizando concretamente a Andalucía de forma diferenciada del resto.
- *Nivel educativo*. Es una variable cualitativa, nominal, categórica (35 categorías). Recoge la variable original del nivel de estudios, facilitada por el Padrón y se convierte en una variable agrupada con cuatro categorías (educación primaria, educación secundaria, bachiller-FP II y universitaria). Conviene destacar la ausencia de información para una parte importante de la muestra y las dificultades de medición de esta característica a partir de la información del Padrón Continuo.
- *Base de cotización diaria.* Variable continua, numérica y de carácter cuantitativo. Esta variable se determina a partir de la base de cotización, siendo ésta una variable original facilitada por la Seguridad Social. Puesto que la base de cotización recoge el importe sobre el que se determinan las cuotas mensuales pagadas a la Seguridad Social y las cantidades a percibir, en su caso, por las prestaciones, se puede emplear como indicador del nivel de renta del autónomo. En concreto la cotización diaria se ha calculado en base al número de días que ha estado vigente la empresa y la base de cotización que ha realizado[9].
- *Ramas de actividad.* Es una variable cualitativa, nominal, categórica (259 categorías). Es una variable original facilitada por la Seguridad Social. En base a la actividad económica de la empresa (cuenta de cotización) que utiliza criterios equivalentes al código CNAE[10] C09C con una codificación a tres dígitos, se

9 Correspondiendo 10.202,40 € a la base mínima en el año 2011, es decir, 850,20 € por las doce anualidades, o a una cotización diaria de 27,95 €.

10 Se utiliza el campo C09C de la tabla de AFILIANON ya que los otros campos que recogen el

ha realizado una agrupación a un dígito con las categorías siguientes: Agricultura, ganadería, caza y pesca; Industria; Construcción; Comercio; Hostelería; Transporte; Educación; Banca y seguros; Administración Pública; Sanidad; Otras actividades.

- *Sector de actividad*. Es una variable cualitativa, nominal, categórica (3 categorías). En base a la actividad económica de la empresa, se obtienen las ramas de actividad, previamente explicadas, y se agrupan a un dígito con las categorías siguientes: sector primario, secundario y terciario.

- **Variable dependiente**

La relación de afiliación con la Seguridad Social tiene una fecha de alta y de baja, la diferencia entre éstas determina la duración de la relación con la Seguridad Social. Esta relación en días se ha distinguido en cuatro categorías según su duración, al igual que se realiza en el informe GEM. La variable que se quiere explicar y predecir a través de las variables independientes ya nombradas, corresponde a la variable *Categoría*, siendo ésta una variable cualitativa, ordinal, con los siguientes niveles:

- *Nacientes <8 meses*
- *Noveles (8-18] meses*
- *Junior (18-42] meses*
- *Consolidadas >42 meses.*

Escoger esta variable como dependiente mostrará cuáles son ls variables que inciden en el inicio del emprendimiento inmigrante y además mostrará de qué manera se relacionan para determinar cuáles de ellas inciden en las diferentes categorías identificadas.

La distribución de inmigrantes emprendedores obtenida para cada categoría es: Naciente (3.168), Novel (1.938), Junior (2.249) y Consolidada (5.606). Se trata, por tanto, de un problema de clasificación multiclase (4 clases), no balanceado y por la naturaleza de la variable dependiente, ordinal.

A partir de la vista minable, se crean dos particiones una para entrenar y otra para validar. Para entrenar el modelo se selecciona una partición de la muestra con el mismo número de patrones de cada clase, 1.450[11] de

CNAE presentan un grado mayor de valores nulos o vacíos.

[11] El tamaño de cada clase para el conjunto de entrenamiento corresponde al 75% de la categoría con un menor número de datos, en nuestro caso la categoría Novel (1.938).

nacientes, noveles, junior y consolidadas. Representan 5.800 patrones correspondientes al 45% del total de la vista minable. De esta forma se convierte un problema de clasificación no balanceado en balanceado y se evita que el modelo clasifique mejor las clases más numerosas. Tal como se indica previamente, para validar el modelo se recogen el resto de patrones no seleccionados, es decir 7.161 casos.

3.4.2. ÁRBOLES DE DECISIÓN

Para resolver problemas de clasificación, existen multitud de técnicas que se pueden utilizar: técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, programación lógica inductiva y métodos basados en funciones de núcleo, entre otros. Cada una de estas técnicas incluye diferentes algoritmos y variaciones de los mismos, así como otro tipo de restricciones que hacen que la efectividad del algoritmo dependa del dominio de aplicación, no existiendo un método universal aplicable a todo tipo de situación (Hernández *et al,* 2004).

De todos los métodos de aprendizaje, los sistemas de aprendizaje basados en árboles de decisión son quizás el método más fácil de interpretar y el más versátil en problemas con diversidad de tipos de variables independientes. Los árboles de decisión son capaces de trabajar tanto con variables continuas como con variables nominales. En nuestro caso, además se dispone de un número alto de variables nominales no ordinales con una alta cardinalidad, teniendo incluso variables continuas.

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera, que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Debido al objetivo de caracterización, explicado previamente, se seleccionó esta metodología primando su interpretabilidad sin dejar de lado su precisión.

El modelo seleccionado C4.5 (Quinlan 1993) es una evolución del modelo de partición ID3 (Quinlan 1983), (Quinlan 1986), ambos modelos están basados en criterios de partición derivados de la ganancia de información. Tienen poda basada en reglas u otros mecanismos más sofisticados. Contiene métodos de

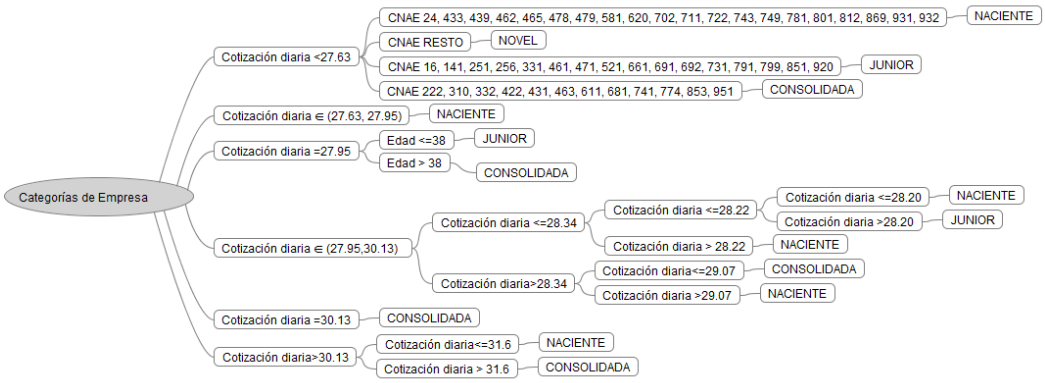
colapsado de ramas y muchas otras mejoras. La versión que se ha utilizado es la J4.8, que se distribuye en la librería WEKA, presentando esta última una versión más avanzada.

Para evitar sobreentrenamiento se han considerado 40 patrones, como nº máximo de patrones por hoja del árbol.

4. RESULTADOS OBTENIDOS

El árbol de clasificación obtenido, muestra que fundamentalmente para identificar la categoría de la empresa, se debe tener en cuenta la Cotización diaria, edad y CNAE.

Gráfico 1. Árbol de decisión



Fuente: Elaboración propia basada en MCVL 2011

El árbol de decisión presenta el siguiente resultado en el porcentaje de clasificación correcta para el conjunto de TEST (70,10%).

Tabla 3. Matriz de confusión

Clasificada como --->	*Naciente*	*Novel*	*Junior*	*Consolidada*
Naciente	1.561	93	35	29
Novel	124	112	130	122
Junior	28	91	341	339
Consolidada	83	182	885	3.006

Fuente: Elaboración propia a partir del modelo obtenido

A la luz de la matriz de confusión (tabla 3), se puede observar que las clases extremas (Nacientes y Consolidada) muestran un porcentaje de clasificación alto. La clase Consolidada, se confunde ligeramente con la clase Junior, debido que el etiquetado de la clase que se ha realizado teniendo en cuenta una frontera numérica (>42 meses). Igualmente se puede ver en el caso de Naciente y Novel. Por último, la clase Novel, no se clasifica correctamente debido, seguramente, a que es una clase central y no presenta diferencias entre Naciente-Junior, con las variables independientes que se utilizan en el modelo.

Tabla 4. Porcentaje de aciertos y errores para cada clase

	Aciertos	Fallos
Naciente	90,86%	9,14%
Novel	22,95%	77,05%
Junior	42,68%	57,32%
Consolidada	72,33%	27,67%

Fuente: Elaboración propia a partir del modelo obtenido

5. CONCLUSIONES

La primera cuestión que conviene destacar es la gran potencialidad de análisis que proporciona la MCVL para complementar otras fuentes de información consagradas como la Encuesta de Población Activa a la hora de conocer las características del colectivo que interesa, en este caso, los inmigrantes que desarrollan una actividad económica por cuenta propia cotizando en el régimen de autónomos de la Seguridad Social. A pesar de su carácter de registro administrativo y de las diversas limitaciones de los datos obtenidos tanto en la naturaleza de la información como en las dificultades de tratamiento técnico, es indudable la fuerza de una muestra tan extensa que permite alcanzar niveles de desagregación y precisión inconcebibles en otras fuentes.

Por otro lado, se ha presentado un modelo para caracterizar las categorías de las empresas; en relación con este modelo, se puede observar:

- Las principales variables son la cotización diaria, el CNAE, y la edad.
- Los resultados que arroja el árbol de decisión señalan que serán

empresas nacientes aquellas en las que la cotización es menor o igual a 27,63 € y su rama de actividad se encuentra entre las siguientes: servicios de apoyo a la silvicultura, acabado de edificios, otras actividades de construcción especializada, comercio al por mayor de equipos para las tecnologías de la información y las comunicaciones, comercio al por menor en puestos de venta y en mercadillos, comercio al por menor no realizado ni en establecimientos, ni en puestos de venta ni en mercadillos, edición de libros, periódicos y otras actividades editoriales, programación, consultoría y otras actividades relacionadas con la informática, actividades de consultoría de gestión empresarial, servicios técnicos de arquitectura e ingeniería y otras actividades relacionadas con el asesoramiento técnico, investigación y desarrollo experimental en ciencias sociales y humanidades, actividades de traducción e interpretación, actividades de las agencias de colocación, actividades de seguridad privada, actividades de limpieza, otras actividades sanitarias, actividades deportivas, actividades recreativas y de entretenimiento.

- Se tratará de empresas junior si la rama de actividad corresponde a cualquiera de los siguientes grupos: actividades de apoyo a la agricultura, a la ganadería y de preparación posterior a la cosecha, confección de prendas de vestir, excepto de peletería, fabricación de elementos metálicos para la construcción, tratamiento y revestimiento de metales; ingeniería mecánica por cuenta de terceros, reparación de productos metálicos, maquinaria y equipo. intermediarios del comercio, comercio al por menor en establecimientos no especializados, depósito y almacenamiento, actividades auxiliares a los servicios financieros, excepto seguros y fondos de pensiones, actividades jurídicas, actividades de contabilidad, teneduría de libros, auditoría y asesoría fiscal, publicidad, actividades de agencias de viajes y operadores turísticos, otros servicios de reservas y actividades relacionadas con los mismos, educación preprimaria, actividades de juegos de azar y apuestas.
- Serán empresas consolidadas aquellas que pertenecen a: fabricación de productos de plástico, fabricación de muebles, instalación de máquinas y equipos industriales, construcción de redes, demolición y preparación

de terrenos, comercio al por mayor de productos alimenticios, bebidas y tabaco, telecomunicaciones por cable, compraventa de bienes inmobiliarios por cuenta propia, actividades de diseño especializado, arrendamiento de la propiedad intelectual y productos similares, excepto trabajos protegidos por los derechos de autor, educación preprimaria, reparación de ordenadores y equipos de comunicación.

- Por último, de acuerdo con los resultados del árbol de decisión, serán empresas nacientes si su cotización se encuentra entre 27.63 y 27,95 donde discrimina de acuerdo con el CNAE tal y como se ha comentado previamente. Además son nacientes las empresas que se encuentran entre 27.95 y 28,20 incluyendo las que son iguales a 28,20. También serían empresas nacientes las que se encuentran entre 28.22 y 28,34 y por último las que su cotización se encuentra entre 30,13 y 31,6. Serán empresas noveles si su cotización es menor a 27,95 de acuerdo con su CNAE, tal y como se ha explicado previamente. Su categoría será junior en dos casos, el primero es si su cotización se encuentra entre 28,20 y 28,22 y el segundo es si la cotización es igual a 27,95 y el empresario tiene menos o 38 años. Y por último serán empresas consolidadas si la cotización es igual a 27,95 y el empresario es mayor de 38 años; también en los casos en que la cotización se encuentra entre 28,34 y 29,07 incluida esta última cifra, y finalmente si la cotización es igual a 30,13, o tiene cotizaciones superiores a 31,6.
- No se clasifica correctamente la clase Novel debido principalmente a que esta categoría de empresa es una variable que identifica el informe GEM, para cuyo análisis no se dispone de suficientes variables que permitan un mayor grado de explicación.

6. AGRADECIMIENTOS

Este trabajo ha sido realizado gracias a los datos facilitados por el Ministerio de Empleo y Seguridad Social sin los cuales no hubiera sido posible el análisis de datos reales en un colectivo tan específico, como son los inmigrantes emprendedores.

Este trabajo ha sido parcialmente financiado por el proyecto TIN2011-22794 del Ministerio de Innovación Ciencia y Tecnología (MICYT), y de los fondos FEDER con el proyecto P2011-TIC-7508 de la Junta de Andalucía.

7. BIBLIOGRAFIA

- ARGIMÓN, I. AND GONZÁLEZ C. I. La muestra continua de vidas laborales de la seguridad social. Boletín Económico. Banco de España 5 (2006): 39-53.
- ESPAÑA, M. T. I. N. Gabinete de Prensa, en el Buscador de noticias del Ministerio, Sección Laboral,[en línea]. (2009).
- HERNÁNDEZ LÓPEZ, M. Predicción económica con algoritmos genéticos: operadores genéricos versus matriz de transición. Estadística Española 46.157 (2004): 389-407.
- HERNÁNDEZ, J., RAMÍREZ, M. Y FERRI, C. (2004). Introducción a la Minería de Datos. Prentice Hall, Madrid.
- LAPUERTA, I. (2010) Claves para el trabajo con la Muestra Continua de Vidas Laborales..
- MORTIMER, J. T. AND SHANAHAN, M. J. eds. (2003) Handbook of the life course. Springer, New York.
- OECD (2012) OECD Territorial Review of Smäland-Blekinge (Sweden), OECD Press: Paris.
- ORALLO HERNÁNDEZ, J., RAMÍREZ QUINTANA, M. J. AND FERRI RAMÍREZ, C.. (2004) Introducción a la Minería de Datos. Pearson Prentice Hall, Madrid.
- PAJARES, M. (2009). Inmigración y mercado laboral. Informe 2009.
- PÉREZ GARCÍA, J. I. (2008). La muestra continua de vidas laborales: una guía de uso para el análisis de transiciones. Revista de Economía Aplicada 16.1: 5-28.
- QUINLAN, J. R. (1983) Learning efficient classification procedures and their application to chess end games. Machine learning. Springer Berlin Heidelberg, pp. 463-482.
- QUINLAN, J. R. (1993) C4. 5: programs for machine learning. Vol. 1. Morgan Kaufmann.
- TOHARIA, L. (2006) Los mercados de trabajo transicionales. Madrid: MTAS.